Expansion of disease gene families by whole genome duplication in early vertebrates

Param Priya Singh

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Param Priya SINGH

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Expansion des familles de gènes impliquées dans des maladies par duplication du génome chez les premiers vertébrés

(Expansion of disease gene families by whole genome duplication in early vertebrates)

Soutenue le **11 Décembre 2013**

Devant le jury composé de :

M. Hugues ROEST-CROLLIUS	Rapporteur
M. Pierre PONTAROTTI	Rapporteur
M. Lluis QUINTANA-MURCI	Examinateur
M. Gilles FISCHER	Examinateur
Mme. Ilaria CASCONE	Examinateur
M. Hervé ISAMBERT	Directeur de thèse

Contents

Acknowledgements iii

Abbreviations & Definitions ix

I Introduction 1

1 Preamble 3
- 1.1 Résumé de la thèse . . . 3
- 1.2 Thesis summary . . . 5
- 1.3 Organization of the thesis . . . 6
- 1.4 Publications resulted/forthcoming from this thesis . . . 6

2 Evolution by Gene Duplication 7
- 2.1 Mechanisms of gene duplication . . . 7
  - 2.1.1 Unequal crossing over . . . 7
  - 2.1.2 Retroposition . . . 9
  - 2.1.3 Non-homologous mechanisms . . . 10
  - 2.1.4 Whole genome duplication . . . 10

3 Evolutionary Constraints & Retention of Duplicated Genes 13
- 3.1 Neofunctionalization . . . 13
- 3.2 Subfunctionalization . . . 14
- 3.3 Buffering against deleterious mutations . . . 14
- 3.4 Dosage balance hypothesis . . . 15

4 Whole Genome Duplications & Evolution of Vertebrates 17

5 Objectives of This Thesis 21

II Materials & Methods 23

6 Identification of Ohnologs 25
- 6.1 Input genomes, orthologs and paralogs . . . 25
  - 6.1.1 Protein coding genes and their genomic coordinates . . . 25
  - 6.1.2 Orthologs and paralogs . . . 27
- 6.2 Identification of the synteny blocks and anchors . . . 28
- 6.3 Calculation of *P-value* to rule out spurious synteny . . . 30
- 6.4 Identify putative ohnolog pairs . . . 31
- 6.5 Combine *P-value* from anchors . . . 32

6.6 Sample genomes with multiple window sizes . . . 32
6.7 Combine *P-value* from all outgroups . . . 33
6.8 Filter ohnolog pairs to remove false positives . . . 33
6.9 Construction of ohnolog families . . . 33
6.10 Randomization of the human genome . . . 34
6.11 Small Scale Duplicates (SSD) . . . 34
6.12 Ohnologs in the teleost fish genomes: the 3R-WGD . . . 35
6.12.1 The 2R-WGD . . . 35
6.12.2 The 3R-WGD . . . 36
6.13 Development of the OHNOLOGS server . . . 36

7 Collection of Cancer/Disease Genes & Functional Genomic Data 37
7.1 Cancer genes . . . 37
7.1.1 Oncogenes and tumor suppressors . . . 39
7.1.2 "Core" cancer genes . . . 40
7.2 Dominant & recessive disease genes . . . 40
7.3 Haploinsufficient and dominant negative genes . . . 40
7.4 Genes with autoinhibitory protein folds . . . 41
7.5 Genes coding for protein complexes . . . 42
7.5.1 Human protein reference database . . . 42
7.5.2 Comprehensive resource of mammalian protein complexes . . . 42
7.5.3 Gene ontology . . . 43
7.5.4 Census of soluble human protein complexes . . . 43
7.5.5 Permanent complexes . . . 43
7.6 Essential genes . . . 43
7.6.1 Human orthologs of Mouse essential genes . . . 43
7.6.2 Human essential genes from *In-vitro* knock-out experiments . . . 44
7.7 Genes with copy number variations . . . 45
7.8 Expression Level . . . 45
7.9 Disease genes in other vertebrates . . . 45
7.9.1 Mouse . . . 46
7.9.2 Rat . . . 46
7.10 Analysis of ohnologs conservation using Ka/Ks ratios . . . 46

8 Causal Inference Analysis 49
8.1 Mediation Analysis . . . 49
8.1.1 Total, direct & indirect effects . . . 49
8.1.2 Mediation calculations . . . 50
8.1.3 Interpretation of Mediation results . . . 51
8.1.4 Application on genomic properties . . . 52

III Results 53

9 Characterization of Vertebrate Ohnologs 55
9.1 Combining information from Multiple outgroups improves ohnolog detection . . . 55
9.1.1 Comparison with randomized human genome . . . 56
9.2 Comparison of ohnologs with published datasets . . . 58
9.3 Ohnolog family size distribution . . . 61

9.4 Ohnolog pairs for other vertebrates . . . 62
9.5 The OHNOLOGS server . . . 63
9.5.1 Search . . . 63
9.5.2 Interpretation of an ohnolog family . . . 64
9.5.3 Browse & Download . . . 65
9.6 Ohnologs in the Teleost fish genomes . . . 66
9.6.1 Ohnologs from the 2R-WGD . . . 66
9.6.2 Ohnologs from the 3R-WGD . . . 67

10 Enhanced Retention of "Dangerous" Genes by WGD 69
10.1 The Majority of "dangerous" genes retain more ohnologs . . . 69
10.1.1 Ohnolog–disease association is consistent for high confidence ohnolog datasets 72
10.1.2 Enhanced retention of "dangerous" ohnologs in Mouse & Rat genomes . . . 73
10.2 "Dangerous" genes show no biased retention by SSD or CNV . . . 74
10.2.1 Small scale duplicates from Ensembl . . . 74
10.2.2 Small scale duplicates from sequence comparisons . . . 76
10.2.3 Ohnolog and SSD retention bias in different human primary tumors . . . 76
10.3 Mapping cancer and disease gene duplications on Ensembl duplication nodes . . . 77
10.4 Ohnologs are more conserved than non-ohnologs . . . 78
10.5 Dominant, and not recessive disease genes have retained more ohnologs . . . 80
10.5.1 Recessive disease genes . . . 80
10.5.2 Essential genes . . . 81

11 Dosage Balance, Expression level & Human Ohnologs 83
11.1 Mixed susceptibility of human ohnologs to dosage balance . . . 84
11.1.1 High retention of protein complexes in ohnologs . . . 84
11.1.2 Transient *versus* permanent complexes . . . 85
11.1.3 Susceptibility of human protein complexes to disease mutations . . . 86
11.2 Gene expression level and human ohnologs . . . 86
11.3 Sequence conservation and ohnolog retention . . . 88

12 Indirect Causes of Ohnolog Retention 91
12.1 The effect of dosage balance is mediated by mutation susceptibility . . . 92
12.1.1 Mediation of 'Dosage.Bal.' $\Longrightarrow$ 'Ohnolog' by 'Delet.Mut.' genes . . . 95
12.1.2 Mediation of 'Delet.Mut.' $\Longrightarrow$ 'Ohnolog' by 'Dosage.Bal.' genes . . . 96
12.1.3 Mediation of 'Dosage.Bal.' $\Longrightarrow$ 'Ohnolog' by 'Delet.Mut.' genes after excluding SSD and CNV genes . . . 97
12.1.4 Mediation of 'Delet.Mut.' $\Longrightarrow$ 'Ohnolog' by 'Dosage.Bal.' genes after excluding SSD and CNV genes . . . 98
12.2 Small effect of essentiality on ohnolog retention . . . 99
12.3 Negative causal effect of high expression on ohnolog retention . . . 100
12.4 Sequence conservation & ohnolog retention . . . 101
12.4.1 Mediation with low Ka/Ks values . . . 101
12.4.2 Mediation with high Ka/Ks values . . . 102

13 Population Genetic Model for the Retention of "Dangerous" Ohnologs 105

IV Discussion & Perspectives 109

14 Discussion & Perspectives 111

A Articles 117

List of Figures 140

List of Tables 141

Bibliography 156

Abbreviations & Definitions

WGD	Whole Genome Duplication
Ohnolog(s)	Gene(s) retained from Whole Genome Duplications
2R-WGD	Two Rounds of Whole Genome Duplication in Vertebrate Ancestor
3R-WGD	Third round of Whole Genome Duplication in bony fishes
SSD	Small Scale Duplication
CNV	Copy Number Variation
Ka	Non-Synonymous Substitution (Also D_N)
Ks	Synonymous Substitution (Also D_S)
MY	Million Years

Part I

Introduction

1 Preamble

1.1 Résumé de la thèse

L'expansion, au cours de l'évolution des vertébrés, de familles de gènes impliqués dans le développement de cancers ou d'autres maladies génétiques chez l'homme, pose question. En effet, alors que ces gènes subissent une forte pression de sélection négative, on constate que leur famille s'est considérablement élargie par le biais de duplications. Dans ce travail, nous avons montré que les copies de ces gènes chez les vertébrés n'ont pas été retenues suite à une duplication locale, mais une duplication globale de génome. En fait, l'expansion de ces familles de gènes est la conséquence de deux duplications globales du génome, chacune couplée à un phénomène de spéciation, qu'a connu l'ancêtre des vertébrés il y a environ 500 millions d'années. Nous avons également montré que la rétention de ces gènes dupliqués par duplication globale de génome — aussi appelé 'ohnologues' — est favorisée par leur susceptibilité aux mutations délétères ayant un phénotype dominant. Par ailleurs, cette rétention se révèle être plus fortement associée à leur susceptibilité aux mutations délétères dominantes, c'est-à-dire à leur "dangerosité", plutôt qu'à leur nature essentielle vis-à-vis de l'organisme.

De nombreuses hypothèses ont été proposées pour expliquer la rétention de copies de gènes après des événements de duplications locales ou globales telles que le niveau d'expression des gènes ou la conservation des séquences. En particulier, la nécessité de maintenir un équilibre entre les différents niveaux d'expression de gènes, ou hypothèse d'équilibre de dosage, a souvent été invoquée comme cause principale de rétention des ohnologues. Afin de distinguer les effets directs et indirects des différents hypothèses proposées, nous avons effectué des analyses statistiques d'inférence bayésienne, ou analyse de médiation. Ces analyses nous ont permis de démontrer que la susceptibilité aux mutations délétères dominantes est en fait la cause directe de rétention des ohnologues dans le génome humain, les autres propriétés ayant un effet indirect sur leur rétention.

Nos résultats nous ont permis d'établir le mécanisme de rétention des copies de gènes après duplication globale de génome. Ce mécanisme repose sur le phénomène de spéciation se produisant après une duplication globale et sur la susceptibilité de certains gènes aux mutations délétères dominantes. Notre équipe a par ailleurs confirmé ce mécanisme grâce à des simulations de ce nouveau modèle de rétention.

De nombreux organismes sont aujourd'hui reconnus comme descendants d'ancêtres polyploïdes ayant survécus à des événements de duplications globales. Ces phénomènes de duplications globales ont joué un rôle majeur dans l'histoire évolutive de ces organismes en offrant d'innombrables possibilité d'innovations. Toutefois, les réarrangements génomiques, la perte de certains gènes et la divergence des séquences rendent difficile l'identification des ohnologues, notamment dans les lignées pour lesquelles les duplications globales sont très anciennes. Dans ce travail, nous avons développé une approche permettant d'identifier efficacement les ohnologues issus de duplications globales. Plus particulièrement, cette approche nous a permis de déterminer les ohnologues issus des deux duplications de génome ayant eu lieu chez l'ancêtre commun à tous les vertébrés, ainsi que les ohnologues provenant des duplications globales chez les poissons. Notre approche exploite les données de génomes complets aujourd'hui disponibles pour certains vertébrés et invertébrés. Une point important de notre algorithme est qu'il nous permet d'établir un indice de confiance (p-value) pour chaque paire d'ohnologues identifiée, basés sur la distribution des homologues dans les génomes comparés.

Nos résultats mettent en avant le rôle important de la sélection négative couplée aux duplications globales de génome dans l'émergence de la complexité des vertébrés, tout en expliquant la surprenante expansion au cours de l'évolution des familles de gènes impliqués dans les maladies génétiques chez l'homme. Par ailleurs, notre algorithme, en permettant une identification des ohnologues avec un indice de confiance, fournit une meilleure compréhension de l'histoire évolutive des génomes, et permet la mise en évidence de régions du génome potentiellement dangereuses à l'échelle de l'individu.

1.2 Thesis summary

The emergence and evolutionary expansion of gene families implicated in cancers and other severe genetic diseases is an evolutionary oddity from a natural selection perspective. Disease genes have been shown to be under strong purifying selection in the human genome. Yet, these gene families have been greatly expanded by duplications in the course of vertebrate evolution, compared to other vertebrate genes without known deleterious mutations. In this work, we have shown that the majority of these genes have not been duplicated through small scale duplications (SSD). Instead, the expansion of these gene families can be traced back to two rounds of whole-genome duplication (WGD), that occurred at the onset of jawed vertebrates, some 500 MY ago. We also observed that the retention of these WGD-duplicated genes — so-called ohnologs — is most enhanced for the genes prone to dominant deleterious mutations and not recessive mutations. This retention was also found to be more strongly associated with their susceptibility to deleterious mutations, than their functional importance in terms of essentiality.

It has been well established that different duplication mechanisms (SSD or WGD) lead to the biased expansion of different categories of genes. Constraint to maintain balanced expression levels or the so called dosage balance hypothesis has been argued to be the major underlying mechanism behind the enhanced retention of ohnologs. To unravel the causal mechanism beyond statistical correlations we used a Mediation analysis in the context of ohnolog retention. Using Mediation analysis, we have shown that the susceptibility to deleterious mutations is a likely direct cause of the retention of ohnologs in the human genome. Furthermore the observed effect of dosage balance constraints, and many other functional properties known to be associated with ohnolog retention, are in fact indirectly mediated by the susceptibility to dominant deleterious mutations.

We have also developed a population genetic model to explain our observations. According to our model, this enhanced retention of ohnologs prone to autosomal-dominant deleterious mutations is a consequence of WGD-induced speciation and the ensuing purifying selection in post-WGD species.

Diverse organisms are now known to descend from polyploid ancestors, often with multiple rounds of WGDs. These ancient polyploidy events are of immense significance in the evolutionary history of organisms, as they are known to facilitate unique evolutionary innovations. However, genome rearrangements, high gene loss, sequence divergence and discrepancies in genome annotation make identification of ohnologs in the extant genomes difficult, especially for old genome duplications. In the present work, we have developed an efficient algorithmic approach to identify ohnologs from the vertebrate ancestral WGD and the fish specific WGD with high confidence. Our approach takes advantage of the availability of multiple complete genome sequences for vertebrates and invertebrates to overcome these challenges. More importantly, we have developed an approach to calculate a confidence index (*P-value*) associated with individual ohnolog pairs based on the conservation of synteny in multiple genomes.

Our findings highlight the importance of WGD-induced non-adaptive selection for the emergence of vertebrate complexity, while rationalizing, from an evolutionary perspective, the expansion of gene families frequently implicated in genetic disorders and cancers. The high confidence ohnologs identified by our approach will pave the way for further analyses in a variety of vertebrate genomes.

1.3 Organization of the thesis

This manuscript has been organized in 4 parts and 14 chapters. Part-I having the first five chapters, including the current one, concerns the introduction of the subject matter of this thesis. In chapter 2 and 3, I attempt to introduce the mechanisms that give rise to duplication of genes and entire genomes, followed by the evolutionary constraints that underlie the retention of only a subset of duplicated genes.

Chapter 4 is focussed on the two round of whole genome duplications in the vertebrate ancestor and the overall impact of these events on the evolution of vertebrates. I will also discuss here, the evolution after whole genome duplications and the approaches used to identify genome duplication in vertebrates, and the challenges in identification of ohnologs. In the last chapter of introduction, I will summarize the questions and problems addressed in this thesis.

Part-II contains the methods used in this work. In chapter 6, I will detail our approach to identify ohnologs from the vertebrate and fish specific WGD. Chapter 7 contains details of collection of data on disease and cancer mutations, and other functional/sequence properties studied in this work. The last chapter summarizes the approach to perform causal inference analysis.

The third part contains the results acquired. In chapter 9, the results of our approach to identify ohnologs have been summarized. Followed by the analysis of these ohnologs in the context of disease mutations and other associated properties in chapter 10 and 11. Chapter 12 explains the results of causal inference analysis, followed by the population genetic model to explain the retention of ohnologs in chapter 13.

Part-IV consists of the last chapter containing general discussion and perspectives followed by lists of table/figure and bibliography.

1.4 Publications resulted/forthcoming from this thesis

(1) **Singh PP***, Affeldt S*, Cascone I, Selimoglu R, Camonis J, Isambert H. (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Reports* Nov. 29;2(5):1387-98.

(2) Affeldt S*, **Singh PP***, Cascone I, Selimoglu R, Camonis J, Isambert H. (2013) Évolution et cancer: Expansion des familles de gènes dangereux par duplication du génome. *Médecine/Science* 4(29): 358-361.

(3) **Singh PP**, Affeldt S & Isambert S. Human dominant disease genes are enriched in paralogs originating from whole genome duplication, *a comment on Chen et al. PLoS Comput Biol. (2013). 9(5):e1003073*. [Submitted to *PLoS Computational Biology*]

(4) Malaguti G **Singh PP** & Isambert S. (2013) On the Retention of Gene Duplicates Prone to Dominant Deleterious Mutations. [Submitted to *Theoretical Population Biology*]

(5) **Singh PP**, Arora J & Isambert H. Improved ohnolog detection using combined synteny information from multiple outgroups [Under Preparation]

Natural selection merely modified, while redundancy created.

– Susumu Ohno

2 Evolution by Gene Duplication

ALTHOUGH duplication of genomic segments has been documented by early geneticists [Sturtevant, 1925; Haldane, 1932; Bridges, 1936], it was not until late 1960's that the potential of gene duplication as an important evolutionary mechanism has been fully appreciated. Genome sequences and sophisticated analysis methods were not available. Yet, based on early experimental observations and limited genomic information Susumu Ohno [Ohno et al., 1968] and Masatoshi Nei [Nei, 1969] highlighted the importance of gene duplication as the primary evolutionary force to create new genes. Some of the initial concepts and theoretical framework of the evolution after gene duplication was laid down by Susumu Ohno in his seminal book, *Evolution by gene duplication* [Ohno, 1970]. We have come a long way since then, and gene duplication has now been firmly established as the primary mechanism of the formation of new genes. In the introduction of this thesis, I will discuss the mechanisms of formation of new genes, evolutionary constraints underlying their retention (or loss), and the overall impact of gene duplications on the evolution of organisms, in particular vertebrates.

2.1 Mechanisms of gene duplication

The widespread occurrence of gene duplication has become apparent with the advent of genome sequencing. A large number of genes in every sequenced eukaryotic genome have considerable sequence similarity and are clearly the products of gene duplication. An additional copy of a gene can be generated by several mechanisms.

2.1.1 Unequal crossing over

The first observation of duplication by unequal crossing over was made by Taylor et al. studying the replication of DNA on individual chromosomes [Taylor et al., 1957]. Crossing over is the exchange of DNA between the two homologous chromosomes e.g. the maternal and paternal chromosome during meiosis. During this process, the homologous regions on the two aligned chromosomes break and then reconnect to create variations by double stranded breaks. However, if the chromosomes are misaligned, this may result in a duplication of the genomic segment on one chromosome and a deletion in the other (Figure 2.1), also called as non-allelic

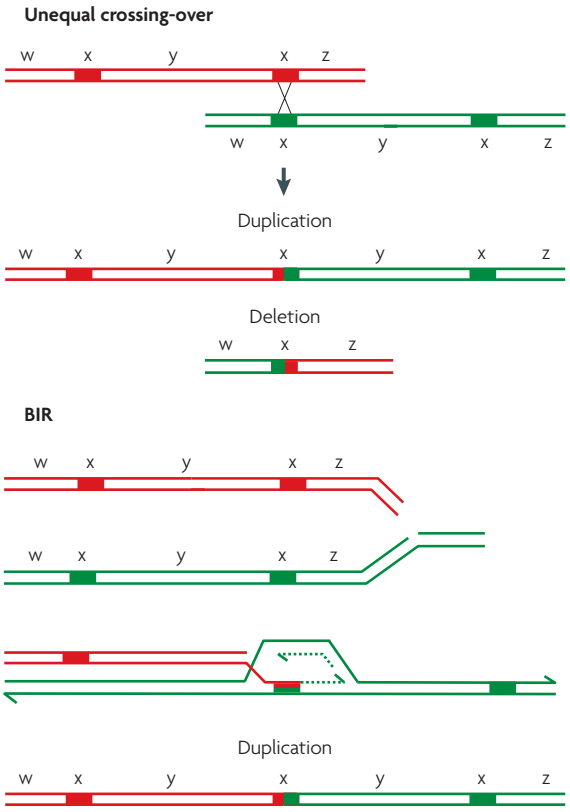

Figure 2.1: Molecular mechanisms of gene duplication by unequal crossing over, or break induced repair (adapted from [Hastings et al., 2009])

homologous recombination. Such a recombination can also occur by break induced repair (BIR). In BIR, broken molecules use ectopic homology to restart the replication fork, which may result in duplication or deletion in distinct events [Hastings et al., 2009].

The gamet with the duplicated segment, when transferred to the next generation results in an organism with the duplicated segment. The duplication (or deletion) of DNA can also occur in the cells during mitotic recombination [LaFave and Sekelsky, 2009]. An unequal exchange of genomic DNA leading to a duplication can also occur between non-homologous chromosomes if there are highly similar sequences on the non-homologous chromosomes.

In the human genome the unequal crossing over can be facilitated by transposable elements such as Alu elements [Hurles, 2004]. Alu elements are 300-base pair dispersed repetitive elements, with approximately 1 million copies in the human genome. Alu element constitute approximately 10% of the entire human genome. Due to their highly similar sequences and high frequency of occurrence they are implicated in unequal crossing over in most of the recent duplications in the human genome [Bailey et al., 2003]. In fact, the rate of unequal crossing over is expected to increase with the amount of repetitive elements in the genomes. The repetitive elements therefore, in part underlie the higher duplicates observed in the metazoan genomes.

A duplication resulted from unequal crossing over can encompass a single gene to a large segment of the genome. If the exchange occurred between the two homologous chromosomes, the duplicated segments would be arranged in tandem. However, in the cases of non-homologous recombination, they may reside on separate unlinked genomic loci. Furthermore, such duplicates tend to preserve the structure of the original gene e.g. exons, introns and regulatory

elements.

The exact contribution of unequal crossing over in generation of small scale and segmental duplication is unclear, however, the analysis of pseudogenes in the human genome have estimated that 20 to 28% of all the identifiable pseudogenes were generated by segmental duplications through unequal crossing over [Torrents et al., 2003; Pei et al., 2012].

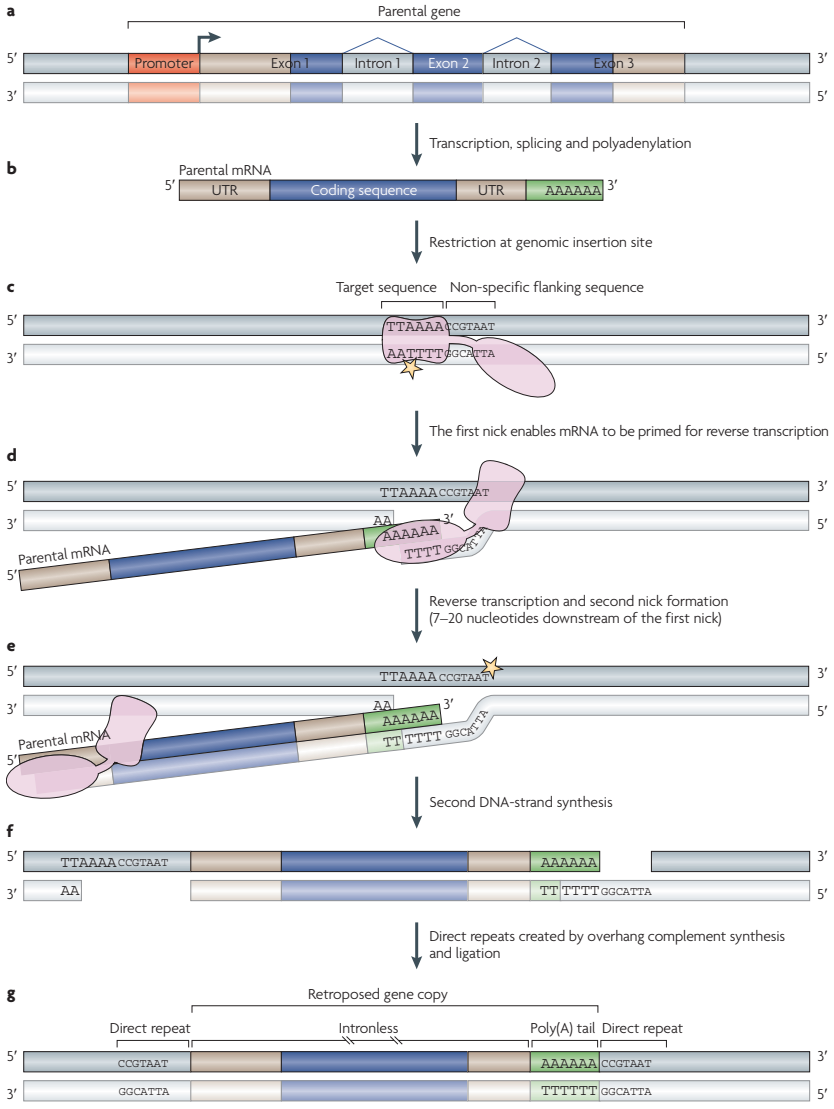

Figure 2.2: Gene duplication by retroposition. Retroposition is initiated with the regular transcription producing a mature mRNA (a-b). At the site of insertion, the endonuclease domain (pink rectangle), creates a first nick (yellow star) (c), this enables the mRNA to be primed for reverse transcription (d-e). A second-strand nick then is generated, leading the cDNA to be inserted at the site flanked by the target sequence (f-g). (adapted from [Kaessmann et al., 2009])

2.1.2 Retroposition

Duplicated copies of genes can also be generated by transposable elements, particularly by retrotransposons. Retrotransposons encode a reveres transcriptase with the endonuclease activity. These reverse transcriptases recognize polyadenylated mRNAs and can reverse transcribe it to DNA. The endonuclease activity leads to the insertion of this reverse transcribed

DNA at a random location in the genome (Figure 2.2). In mammals, LINE-1 or L1 elements (Long Interspersed Elements) are responsible for gene duplication by retroposition [Moran et al., 1999]. Since the retroposition process requires the expression of mRNA, the genes that are expressed in germline are the most likely candidates of retroposition [Kaessmann et al., 2009].

The DNA that is inserted in the genome by this process comes from the processed mRNAs, therefore, such genes lack regulatory elements, introns and have long poly-A tails. For these genes to be functional they must recruit a promoter or other regulatory elements at their new location from the nearby genes. It has been thought earlier that retroposition only leads to non-functional processed cDNA copies, however, recent genome wide studies have shown that these genes are often functional [Kaessmann et al., 2009]. Yet, since their success depends on the recruitment of regulatory elements, a large number of pseudogenes in the human genome (> 80%) correspond to 'processed' pseudogenes generated by retroposition[Pei et al., 2012].

2.1.3 Non-homologous mechanisms

Apart from homologous mechanisms, many non-homologous mechanisms can also lead to duplication of genomic DNA during replication. One such mechanism is the slipping of replication fork [Levinson and Gutman, 1987]. During DNA replication on the lagging strand, if the replication machinery is displaced and subsequently misplaced on another location, it may lead to a duplication or deletion of DNA segment in the genome. Typically, the lengths of such duplicated segments are short (tens to hundred base pairs), however, it can also lead to spontaneous duplication of large genomic regions spanning many genes [Koszul et al., 2003].

In some cases, the lagging strand stalls and subsequently exchanges the template with another replication fork which has a sequence with micro-homology. This may lead to duplication of a part of misaligned segment, and is referred to as Fork Stalling and Template Switching (FoSTeS) [Hastings et al., 2009]. A variation of FoSTeS, called microhomology-mediated break-induced replication (MMBIR) is also known to underlie many structural variations and duplications in the human genome [Zhang et al., 2009; Hastings et al., 2009].

2.1.4 Whole genome duplication

All the above mechanisms generate duplicated regions ranging from a few base pairs to a large genomic segment, typically arranged in tandem. Throughout this thesis, we will refer them as small scale duplicates (SSD). Whole genome duplication (WGD) or polyploidy, however, can give rise to duplication of the entire genome. Such a genome duplication can be achieved by two mechanisms, auto- and allo-polyploidy. Autopolyploidy, can occur by incomplete chromosome segregation, cytokinesis defects or fusion of two cells of the same organism during early development, leading to a polyploid embryo. In case of allopolyploidy, two cells from different but closely related organisms can fuse and give rise to an organism with whole genome duplication.

Traditionally, polyploidy had been considered to be an evolutionary dead end [Mayrose et al., 2011]. Many polyploid plant species were known to early geneticists and it was believed that once an organism becomes polyploid, due to the very nature of the event, the odds of further 'evolution' diminish. It was also believed that animals, unlike plants should not tolerate polyploidy due to the mode of sexual reproduction [Muller, 1925; Mable, 2004]. Susumu Ohno, however proposed that genome duplications are a significant mechanism of evolution even in the animal genomes [Ohno et al., 1968; Ohno, 1970].

Growing genome sequences and state-of-the-art analysis approaches have now established

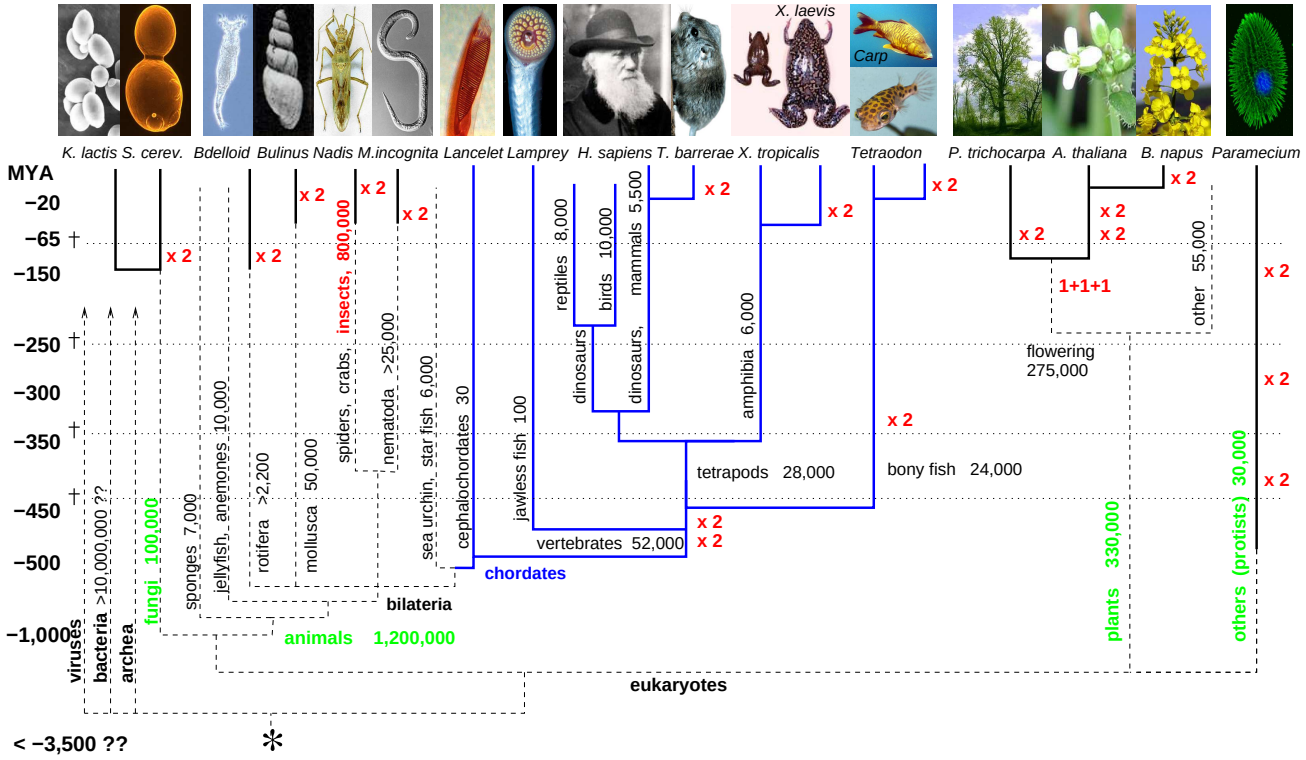

Figure 2.3: Whole genome duplications in the evolution of eukaryotes

polyploidy as a major evolutionary mechanism in all major eukaryotes — from unicellular eukaryotes, fungi plants and animals (Figure 2.3). Polyploidy is especially common in plants (Figure 2.4)[1]. The common ancestor of all the extant angiosperms have undergone a tetraploidy event [Jiao et al., 2011], and almost all major plant lineages have undergone multiple polyploidy events subsequently.

Successive WGDs have also occurred in many animal genomes. Even though most extant species are diploids, polyploid species are also suspected or confirmed in most invertebrate phyla, as in annelids (e.g., leeches [Gambi et al., 1997]), flatworms (e.g., *Stenostomum* [Gregory et al., 2000]), mollusks (e.g., Pacific oyster, *Crassostrea gigas* [Eudeline et al., 2000]) and in the major classes of arthropods, including insects (e.g., *Nabis pallidus* [Grozeva et al., 2004]), maxillopods (e.g., copepods [Gregory et al., 2000]) and branchiopods (e.g., brine shrimp [Dufresne and Hebert, 1995]) (Figure 2.3). Polyploidy in the animal kingdom is specially common in the amphibians e.g. *Xenopus laevis* [Hughes and Hughes, 1993].

Most importantly, most vertebrates are now known to descend from a single lineage that experienced two consecutive WGDs soon after the divergence of other chordates about 500MY ago (this is the long debated "2R hypothesis" [Ohno et al., 1968; Ohno, 1970], Chapter 4). Similarly, all bony fishes, which make up about 90% of the extant fishes, are now known to derive from a single species that doubled its genome about 300MY ago (*i.e.* the "3R hypothesis" [Amores et al., 1998; Jaillon et al., 2004], Chapter 4). In addition, the common carp (*Cyprinus carpio*) and the salmonidae fish (salmon, trout) have also experienced another recent WGD [David et al., 2003]. Hence, there are, for instance, 4 consecutive WGDs between the cephalochordate ((or lancelet) Amphioxus (*Branchiostoma floridae*) and the common carp, with most tetrapods (including mammals) in between at +2WGDs from *Amphioxus* and –2WGDs from

[1] http://genomevolution.org/wiki/index.php/Plant_paleopolyploidy

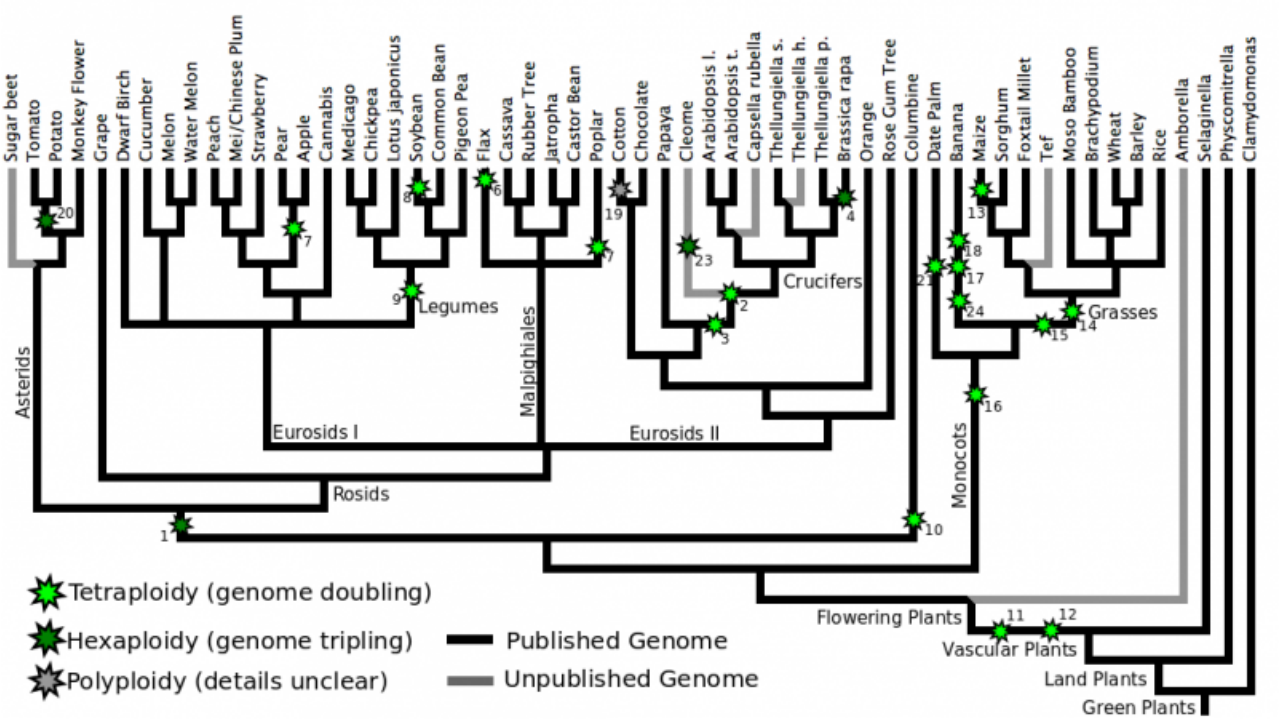

Figure 2.4: Whole genome duplications in the plant kingdom. Figure adapted from CoGEPEdia.

carp and most bony fish at +3WGDs from amphioxus and –1WGD from carp.

Many plant species and animals such as carp, salmon, trout, as well as the amphibian *Xenopus laevis* and the mammal *Tympanoctomys barrerae* (red vizcacha rat from Argentina [Gallardo et al., 2006]) are in fact pseudotetraploid organisms with about twice as many gene loci as their close relatives lacking this recent WGD. Such 'neopolyploid' organisms still carry clear evidence of polyploidy in their genomes.

However, in case of most of the other WGDs, such as the 2R-WGD in vertebrates or the 3R-WGD in bony fishes, the evidence in the genome is not direct. In fact, the genome of these organisms are genetically diploid, yet the evidence that their ancestors underwent WGDs can be sought in the structure of their genomes (Discussed in Chapter 4). Such genetically diploid but structurally polyploid genomes are referred to as paleopolyploid genomes. Before the advent of genome analysis techniques, polyploidy could only be recognized by counting number of chromosomes, or by analyzing the distribution of a handful of sequences on genomes. Hence, many of these paleopolyploidy events have been debated for very long time.

Although in the short term, polyploidy leads to population bottleneck and possible competition with their diploid ancestors, the success of paleopolyploid organisms and the frequent occurrence of a number of neopolyploids strongly suggests that whole genome duplication has been central to the evolution of a vast number of organisms.

3

Evolutionary Constraints & Retention of Duplicated Genes

GENOME wide analyses have estimated the rate of occurrence of gene duplications to be close to nucleotide substitutions, on the order of 0.01 per gene per million years [Lynch and Conery, 2000]. Although the fact that many of the genes are retained by the ancestral WGDs in the studied organisms was not taken into account, the knowledge of the frequent occurrence of gene duplicates has also led to the realizations that not all the duplicates are retained in the genome. Immediately after duplication, both the copies (called paralogs) are expected to perform identical functions. If this redundancy is not advantageous, during subsequent evolution one of the two paralogs becomes typically a non-functional pseudogene by accumulating degenerating mutations and the traces of the duplication are lost from the genome as it continues to harbour mutations without any selective pressure. This stochastic silencing of one of the copy is called as *pseudogenization*, and is considered to be the most likely fate of a duplicated gene. However, in some cases, both duplicated copies are retained in the genome, diverge and become eventually fixed in the population. What evolutionary constraints underlie the retention of only a small fraction of genes? Many mechanisms have been proposed to explain the retention of genes after duplication.

3.1 Neofunctionalization

The idea of neofunctionalization had been proposed very early by Haldane and Fisher [Dittmar and Liberles, 2011], and was followed by Susumu Ohno [Ohno, 1970]. Ohno argued that in the absence of a redundant copy, the function performed by any gene would be under strong selective constraint. However, if an additional copy is available by gene duplication, this purifying selection would be relaxed. While one of the two paralogs provides the ancestral function, the other one would be free to accumulate mutations, and on occasions would evolve a new beneficial function. Both copies would now be retained through positive selection.

However, clear and well studied examples of neofunctionalization are difficult to find [Dittmar and Liberles, 2011; Gibson and Goldberg, 2009]. Under the light that most mutations are either neutral and deleterious, the neofunctionalization concept has been under scrutiny since

its inception. Most claims supporting neofunctionalization have been indirect. For example, on the genome-wide scale the asymmetric divergence pattern of the duplicate copies has been taken to be the evidence of neofunctionalization [Kellis et al., 2004; Byrne and Wolfe, 2007]. Neofunctionalization at the regulatory level has been assessed and supported by the overlapping expression profiles of the duplicated genes [Tirosh and Barkai, 2007; Assis and Bachtrog, 2013]. Acquisition of new domains by the duplicated genes have also been taken to be the evidence for neofunctionalization [Drea et al., 2006; Casewell et al., 2011; Jin et al., 2009]. In most reported instances of neofunctionalization, the neofunctionalized gene performs a function not performed by the ancestral gene, yet the function would exist in some other gene in the genome. Therefore, many instances of proposed neofunctionalization can be accounted for by domain shuffling coupled to fixed duplicates. Evolution of truly 'new' functions, non-existent in the genome before duplication is expected to be very rare.

3.2 Subfunctionalization

Another mechanism that can leads to the fixation of duplicated genes was initially proposed by [Hughes, 1994], and studied extensively by [Force et al., 1999; Lynch and Force, 2000a], known as subfunctionalization. According to this model, both copies continue to accumulate mutations after duplication, and degenerate to a level where none of the paralogs alone can provide old ancestral functions. Hence, the duplicated copies can be preserved for longer times, as the loss of any of the paralog is expected to become deleterious after subfunctionalization. In general, such a subfunctionalization mechanism (also called Duplication Degeneration Complementation or DDC) is neutral and because the duplicates can no longer be lost, population genetic processes ultimately lead to the eventual fixation of the duplicated copies [Lynch and Force, 2000a]. However, several variants of the subfunctionalization model exist and some of the mechanisms are also proposed to be adaptive.

For example, in cases where a single gene performs multiple functions, subfunctionalization may lead to uplifting of such pleotropic constraints, and hence can be beneficial [Hughes, 1994]. The duplicated copies can now escape from the adaptive conflict (the EAC model) of the ancestor and can now become specialized, which was not possible with a single ancestral copy [Des Marais and Rausher, 2008]. A similar model referred to as Adaptive Radiation or Innovation Amplification Divergence (IAD) provides another mechanism of subfunctionalization based on an initial advantageous dosage increase [Francino, 2005; Näsvall et al., 2012]. The boundary between all these mechanisms is not clearly defined and specific evidences in support of each of these models are available. Yet, on a genome wide scale, subfunctionalization appears to be a common mechanism with vast number of gene duplicates with high sequence similarity, related functions and non-overlapping expression profiles [Duarte et al., 2006; de Souza et al., 2005].

3.3 Buffering against deleterious mutations

Several studies have argued that gene duplicates can compensate each other's function and hence, may lead to robustness against deleterious mutations. It had been noticed that gene duplicates confer robustness against null mutations in yeast [Gu et al., 2003; Gu, 2003] and *Caenorhabditis elegans* [Conant and Wagner, 2004]. Such a buffering mechanism requires, however, that the functions of the duplicated copies do not diverge too much.

Functional compensation against deleterious mutations have been also argued by the observations that the disease genes in the human genome have more duplicates as compared to non disease genes. Furthermore, these duplicated genes tend to have higher sequence similarity and co-expression than the duplicates of non-disease genes [Chen et al., 2013b; Dickerson and Robertson, 2011].

Another line of evidence for buffering against deleterious mutations comes from the analysis of essential genes, *i.e.* genes in which silencing or knock-out mutations lead to lethality or sterility. However, the results have been equivocal. If the buffering against deletion for essential genes is indeed beneficial, on a genome wide scale essential genes should have more duplicates than non-essential genes. While some studies indeed observed an enrichment of essential genes in gene duplicates [Gu et al., 2003; Kamath et al., 2003; Makino et al., 2009], many others however have failed to observe any such bias [Liang and Li, 2007; Liao and Zhang, 2007; Guan et al., 2007]. The relationship between gene duplication and functional compensation has remained enigmatic.

3.4 Dosage balance hypothesis

The concept of neo- and subfunctionalization to explain retention of duplicated genes were developed before the distinct properties of the genes retained from small scale duplication (SSD) and whole genome duplication (WGD) became apparent. Therefore, these models do not account for the fact that the genes retained from SSD or WGD have very distinct functional properties. As more genome sequences of paleopolyploid genomes became available it was realized that SSD and WGD-duplicated genes have very distinct functional properties [Davis and Petrov, 2005; Hakes et al., 2007; Fares et al., 2013].

Indeed, WGD-duplicated genes, have been preferentially retained in specific gene classes associated with organismal complexity, such as signal transduction pathways, transcription networks, and developmental genes [Maere et al., 2005; Blomme et al., 2006; Freeling and Thomas, 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010]. By contrast, gene duplicates coming from SSD are strongly biased toward different functional categories, such as antigen processing, immune response, and metabolism [Huminiecki and Heldin, 2010]. SSD and WGD duplicated genes also differ in their gene expression and protein network properties [Hakes et al., 2007; Guan et al., 2007]. Furthermore, recent genome-wide analysis have shown that WGD genes in the human genome have experienced fewer SSD than non-WGD genes and tend to be refractory to copy number variation (CNV) caused by polymorphism of small segmental duplications in human populations [Makino and McLysaght, 2010].

Dosage balance hypothesis has been proposed to explain these antagonist retention patterns of WGD and SSD/CNV gene duplicates. The dosage of a gene *i.e.* the amount of protein expressed has been proposed to play an important role for proper formation or functioning of cellular assemblies such as transcription factors and protein complexes [Birchler et al., 2001; Veitia, 2002, 2003]. Although the early studies focussed on the stoichiometric imbalance of regulatory complexes, [Papp et al., 2003] associated all protein complexes with the dosage balance hypothesis. Studying the yeast duplicates, they observed that only WGD-retained genes are enriched in protein complexes and interpreted that an imbalance in the components of protein complexes leads to lower fitness. [Papp et al., 2003] also set the precedent to seek enrichment of protein complexes in the WGD-retained genes in the support of dosage balance hypothesis.

In general, the dosage balance hypothesis posits that the relative amount of interacting gene products e.g. subunits of protein complexes is crucial for its proper formation and func-

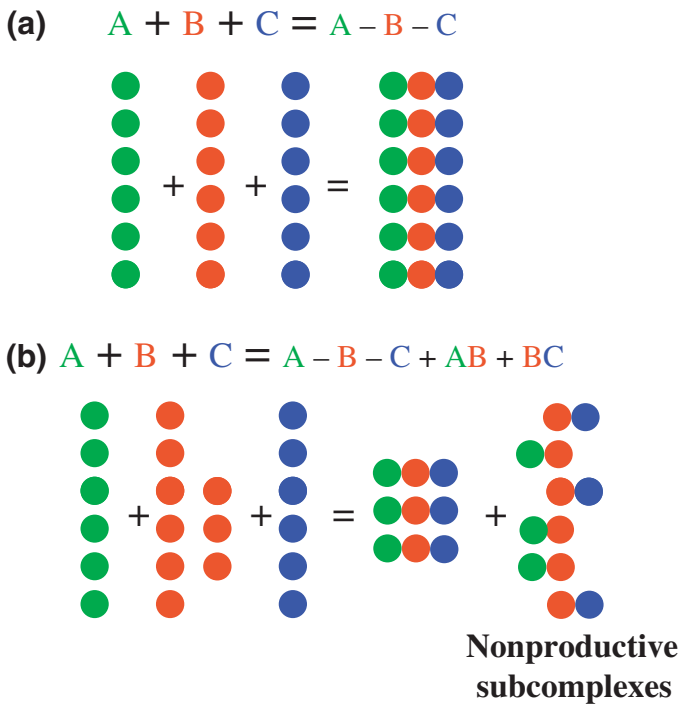

Figure 3.1: Dosage imbalance in macromolecular complexes. (a) A trimer A-B-C is assembled irreversibly from three subunits: A, B and C. A 1:1:1 ratio of each of the subunit leads to proper formation of the trimer. (b) An imbalance in this stoicheometry is caused by the over-expression of one subunit B, that significantly reduces the yield of the functional trimer A-B-C (adapted from [Birchler and Veitia, 2010])

tioning. Duplication of only one of the interacting partner by SSD leads to dosage imbalance such as a stoichiometric imbalance during formation of protein complexes (Figure 3.1). This not only leads to a decreased yield of functional complex, but might also lead to dominant negative or toxic effects by partially assembled non-productive subcomplexes [Papp et al., 2003]. However, in WGD since all the genes are duplicated simultaneously, the relative amount of all the interacting partners is preserved. Yet, just as the duplication of individual genes, the deletion after WGD will also lead to dosage imbalance and hence would be opposed by selection. The dosage balance hypothesis has been frequently invoked to explain the biased retention of SSD and WGD genes in a variety of organisms such as yeast [Papp et al., 2003], Paramecium [Aury et al., 2006], Arabidopsis [Maere et al., 2005] and humans [Makino and McLysaght, 2010], by seeking enrichment of protein complexes in WGD duplicates.

4

Whole Genome Duplications & Evolution of Vertebrates

Duplication of genes and their subsequent divergence has now been established as the major evolutionary mechanism to generate new genes and rewire cellular pathways and networks. While small scale duplications (SSD) provides a continuous flux of genetic material, large scale sporadic duplications of entire genomes are now known to play very important role in the evolution of many organisms. Such whole genome duplication (WGD) events have now been firmly established in almost all major eukaryotic lineages [Van de Peer et al., 2009]. Organisms as diverse as *Saccharomyces cerevisiae*, paramecium, vertebrates and flowering plants are known to be descended from polyploid ancestors, often with multiple rounds of WGDs (Figures 2.3, 2.4). These organisms are referred to as paleopolyploids, as this polyploidy can only be observed at a structural and not at the genetic level in their extant genomes.

In vertebrates in particular, genome duplications were first proposed by Susumu Ohno [Ohno et al., 1968] (the 2R-hypothesis) who argued that the huge leap in complexity in vertebrates was facilitated by sudden burst of increase in genetic material by WGDs, which was tinkered by millions of years of subsequent evolution. The compelling evidence in support of 2R hypothesis remained elusive for a long time, but recent genome wide studies strongly advocate these vertebrate ancestral WGD events [McLysaght et al., 2002; Dehal and Boore, 2005; Putnam et al., 2008]. Due to pioneering works of Susumu Ohno in this area, the genes retained from whole genome duplications are now referred to as *"ohnologs"* [Wolfe, 2000].

These ancient polyploidy events are of immense significance, as they are known to facilitate unique evolutionary innovations. It is now known that ohnologs and SSD contribute antagonistically to different functional categories consistently across diverse taxa, with ohnologs primarily associated with signaling pathways and developmental genes. Although a basic set of developmental genes was already present in chordates, the selective expansion of these gene families in vertebrates by WGDs led to evolution of complex developmental processes [Holland et al., 1994] including neural creast, [Holland et al., 2008] vertebrate skeleton [Wada, 2010], brain [Holland and Short, 2008] and adaptive immune system [Okada and Asai, 2008]. HOX gene clusters involved in patterning of the body plan have also been duplicated primarily by WGD in both vertebrates and teleost fishes [Soshnikova et al., 2013; Kuraku and Meyer,

2009]. Furthermore, gene families involved in many key cellular processes and pathways have retained an excess of ohnologs as compared to SSDs e.g. haemoglobins [Opazo et al., 2012; Storz et al., 2012], ligand-receptor interactions [Braasch et al., 2009], ABC [Annilo et al., 2006], glycolytic pathway [Steinke et al., 2006], GPCRs [Semyonov et al., 2008], GATA transcription factor [Gillis et al., 2009], tetraspanins [Huang et al., 2010], 14-3-3 binding proteins family [Tinti et al., 2012]. These observations have strengthened the hypothesis that WGDs have indeed played a major role in the evolution of vertebrate complexity. Therefore analyzing ancient WGD events, and identification and characterization of ohnologs is central to the understanding of the evolutionary history of paleopolyploid organisms.

Accurate identification of ohnologs from ancient WGD events however, is not straightforward. A typical evolutionary scenario after WGD is presented in Figure 4.1 A-D. Each gene has been labelled by a number and orthologous/paralogous relationships (genes of same color) are depicted by lines. Immediately after WGD all the genes exist in twice as many identical copies preserving their order on chromosomes (Figure 4.1 B). However, this pattern will be faded progressively during the course of evolution primarily by random gene loss from one of the sister regions (Figure 4.1 C). As the polyploid genome eventually becomes genetically diploid again, typically only 10-15% of the genes retain their ohnologs [Brunet et al., 2006]. Furthermore, a combination of inter and intra-chromosomal rearrangements, small scale duplications (not depicted in Figure 4.1 B for simplicity) and sequence divergence of retained paralogs lead to the fractionation of large duplicated regions into smaller degenerated segments with only a few duplicated genes (Figure 4.1 B). This process is termed as rediploidization after WGD.

Such structurally degenerated duplicated regions covering a large proportion of genome are the signals of an ancient polyploidy event and can be identified when the extant paleoployloid genome is compared to itself. If a genome has experienced an ancient WGD event, large blocks of conserved synteny would be detected where all the paralogs duplicated at the suspected time of duplication. For example, using a criteria that at least two consecutive genes duplicated by WGD occur in a window size of three, paralogous synteny blocks can be identified in Figure 4.1 I, highlighted in boxes. The genes on such blocks, if they have been duplicated at the time of WGD event can be characterized as ohnologs.

Another, more convincing evidence of an ancient WGD event can be sought in the comparisons between genome of paleopolyploid organism with an organism diverged before the WGD event, called as an outgroup organism (Figure 4.1 E & G). The outgroup genomes would also undergo independent lineage specific rearrangements (depicted by shuffled location of the genes in Figure 4.1 F & H). Yet each region in the outgroup genome should ideally be paralogous to two regions in the WGD genome, a pattern called doubly conserved synteny (Figure 4.1 J & K). The number of degenerate synteny blocks sharing the same outgroup block can vary depending on the number of WGD experienced by paleopolyploid organism. Such signals covering the majority of the genomes have provided compelling evidence of a WGD in a variety of organisms [Kellis et al., 2004; Jaillon et al., 2004, 2007; Putnam et al., 2008].

However, accurate identification of ohnologs in paleopolyploid genomes poses many challenges, especially in case of old whole genome duplication events such as vertebrate ancestral WGDs. First, only a small fraction of genes retain their ohnolog partners. Second, inter- and intra-chromosomal rearrangements in both outgroup and WGD genome lead to the weakening of synteny conservation making the identification of the synteny blocks difficult. Third, even in the case of retained duplicate partners, sequence divergence and the criteria to identify them (e.g. E-value in BLAST or gene family construction) can lead to spurious orthologs/paralogs assignment, and may also lead to discrepancies in the estimation of their duplication time. Fourth, the levels of genome annotation vary also vastly for different genomes, such as the

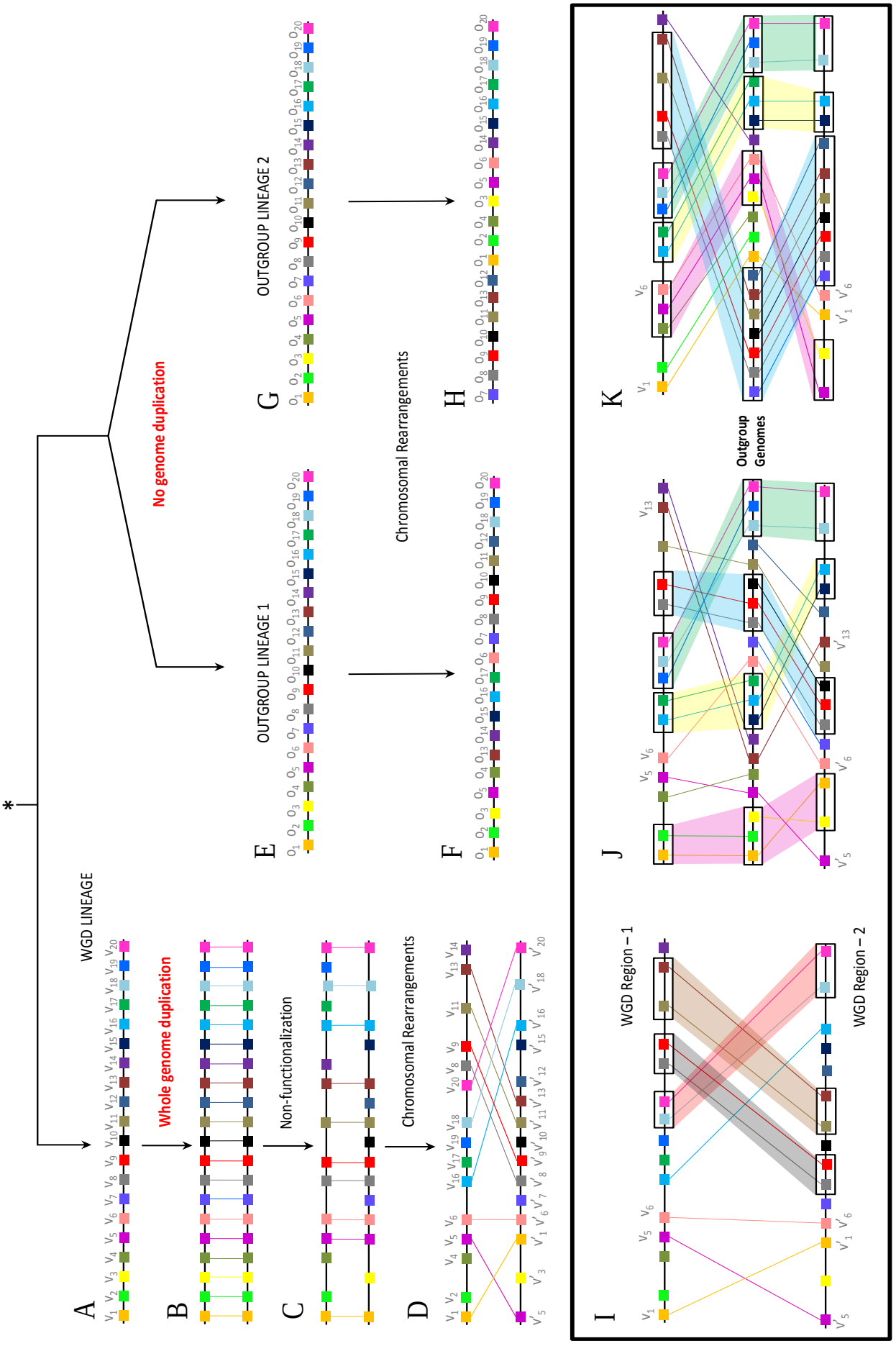

Figure 4.1: Evolution after WGD and identification of ohnologs using self and outgroup synteny comparison

predicted number of genes, their location and the quality of the genome assembly, affecting ohnolog detection. Lastly, window based approaches are typically used to detect such signals; however, in most studies arbitrary window sizes and arbitrary numbers of orthologs/paralogs are used to make synteny calls. Moreover, a quantitative measure to identify the quality of ohnologs has been lacking. Many studies identified human ohnologs by detecting such block within the human genome using a sliding window based approach [Dehal and Boore, 2005; Makino and McLysaght, 2010], yet surprisingly only a few studies have attempted a comparison with a single chordate outgroup genome — *Amphioxus* — with the human genome to identify regions descended from WGDs [Abi-Rached et al., 2002; Putnam et al., 2008].

Due to lineage specific rearrangements it is highly likely that many of the 'true' ohnologs would be missed if the same window size and only a single synteny comparison is performed. For example, note that using a criteria of 2 consecutive candidates in a window of size 3, (Figure 4.1 I-K) many of the ohnologs would not be identified by either self comparison or the outgroup comparison alone e.g. ohnolog pairs $V_1 - V_1^{'}$, $V_5 - V_5^{'}$ are missed as ohnologs by self synteny comparison, but are covered by outgroup-1 and 2 respectively Figure 4.1 I-K. Therefore, it is really important to perform synteny comparisons with multiple outgroups in addition to the genomic regions in the same vertebrate genome. More importantly, as only 10-15% genes are typically retained in synteny after any typical WGD, in outgroup comparison more power in synteny identification can be generated by the orthologs.

5
Objectives of This Thesis

DOMINANT deleterious mutations, that are lethal or drastically reduce fitness over the lifespan of organisms, must also impact their long term evolution on timescales relevant for genome evolution (e.g., >10–100 MY). In fact, dominant disease genes in humans have been shown to be under strong purifying selection. Yet, disease gene families implicated in cancer and severe genetic diseases have also been greatly expanded by duplication in the course of vertebrate evolution. Indeed, considering that many vertebrate disease genes are phylogenetically ancient [Domazet-Lošo and Tautz, 2008; Cai et al., 2009; Dickerson and Robertson, 2011], and that their orthologs also cause severe genetic disorders in extant invertebrates [Berry et al., 1997; Ciocan et al., 2006; Robert, 2010], it is surprising that dangerous gene families have been duplicated more than other vertebrate genes without known dominant deleterious mutations. Furthermore, in invertebrates and vertebrates, very different classes of genes has been expanded by duplications [Putnam et al., 2008]. While gene duplicates might confer mutational robustness against loss-of-function mutations, multiple copies of genes prone to gain-of-function mutations are expected to lead to an overall aggravation of a species' susceptibility to genetic diseases and thus be opposed by purifying selection.

Two alternative hypotheses can be put forward to account for the surprising expansion of dangerous gene families. Either, the propensity of certain genes to acquire dominant deleterious mutations could be a mere by-product of their presumed advantageous functions. In that case, only the overall benefit of gene family expansion should matter, irrespective of the mechanism of gene duplication. Alternatively, gene susceptibility to dominant deleterious mutations could have played a driving role in the striking expansion of dangerous gene families. But what could have been the selection mechanism?

The first aim of this study was to characterize the mechanism of the duplication of cancer and disease genes. Unlike many previous studies, we aimed to generate a comprehensive dataset of Mendelian disease genes and cancer genes, including many other classes of genes known to be susceptible to deleterious mutations. We wanted to investigate the retention pattern of these disease genes after SSD and WGD with the most comprehensive and a high confidence subset of disease genes.

To discriminate between Ohnologs and SSDs, our next aim was to develop an improved method to detect ohnologs in the human genome. In particular, we aimed to use the power of

multiple genome comparisons to overcome many limitations associated with the identification of ohnologs. More importantly, our objective was to calculate a confidence measure to quantify the quality of the identified ohnologs.

We hypothesized, that the susceptibility of these genes must play the central role in their biased retention after duplication in the human genome. A similar proposition had been made in an article comment by Gibson and Spring [Gibson and Spring, 1998]. In fact, multiple genomic properties have been reported to correlate with the retention of genes after WGD or SSD. Therefore, we aimed to go beyond statistical correlations and to use advanced statistical approaches to disentangle the relative effects of mutation susceptibility and other properties, such as dosage balance in particular.

Ultimately, we wanted to put forward a comprehensive model to explain our observations from statistical and advanced inference analysis from a population genetic perspectives.

Part II

Materials & Methods

6
Identification of Ohnologs

THE identification of the genes retained from whole genome duplication is not straightforward. During millions of years of evolution following WGD, as the paleopolyploid genome is reduced to normal ploidy levels, sister regions created by WGD are redistributed across the genome by rearrangements and degenerated by the loss of the majority of ohnologs (Figure 4.1). We used a window based approach to detect such regions, called synteny blocks, between a pair of outgroup (invertebrates) and paleopolyploid (vertebrates) genomes. We compared each vertebrate genome to outgroup genomes (outgroup comparison) and to itself (self-comparison). A summary of our overall approach is detailed in Figure 6.1.

6.1 Input genomes, orthologs and paralogs (6.1A)

To identify ohnologs retained from the 2R-WGD, we used six invertebrate genomes, one lancelet (cephalochordate): Amphioxus (*Branchiostoma floridae*), two tunicates (urochordates): *Ciona intestinalis* and *Ciona savignyi*, an echinoderm: sea urchin (*Strongylocentrus purpuratus*), and two basal bilaterians: fly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) as outgroups. Using these outgroups we identified ohnologs in six completely sequenced vertebrate (tetrapod) genomes: human (*Homo sapiens*), chicken (*Gallus gallus*), dog (*Canis lupus familiaris*), pig (*Sus scrofa*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) (Figure 6.2).

6.1.1 Protein coding genes and their genomic coordinates

We limited ourselves to protein coding genes. Except for sea urchin and Amphioxus, the protein coding genes and their genome positions were obtained from Ensembl version 70 [Flicek et al., 2013] using BioMart. Sea urchin and Amphioxus genes and their genome coordinates were downloaded from Ensembl Metazoa [Kersey et al., 2012] and DOE Joint Genome Institute (JGI) [Putnam et al., 2008] respectively. We further excluded genes belonging to unassembled scaffolds or haplotype regions in the vertebrate genomes. Each outgroup and vertebrate genome was then represented by a list of gene identifiers sorted on the basis of their start positions on their respective chromosome.

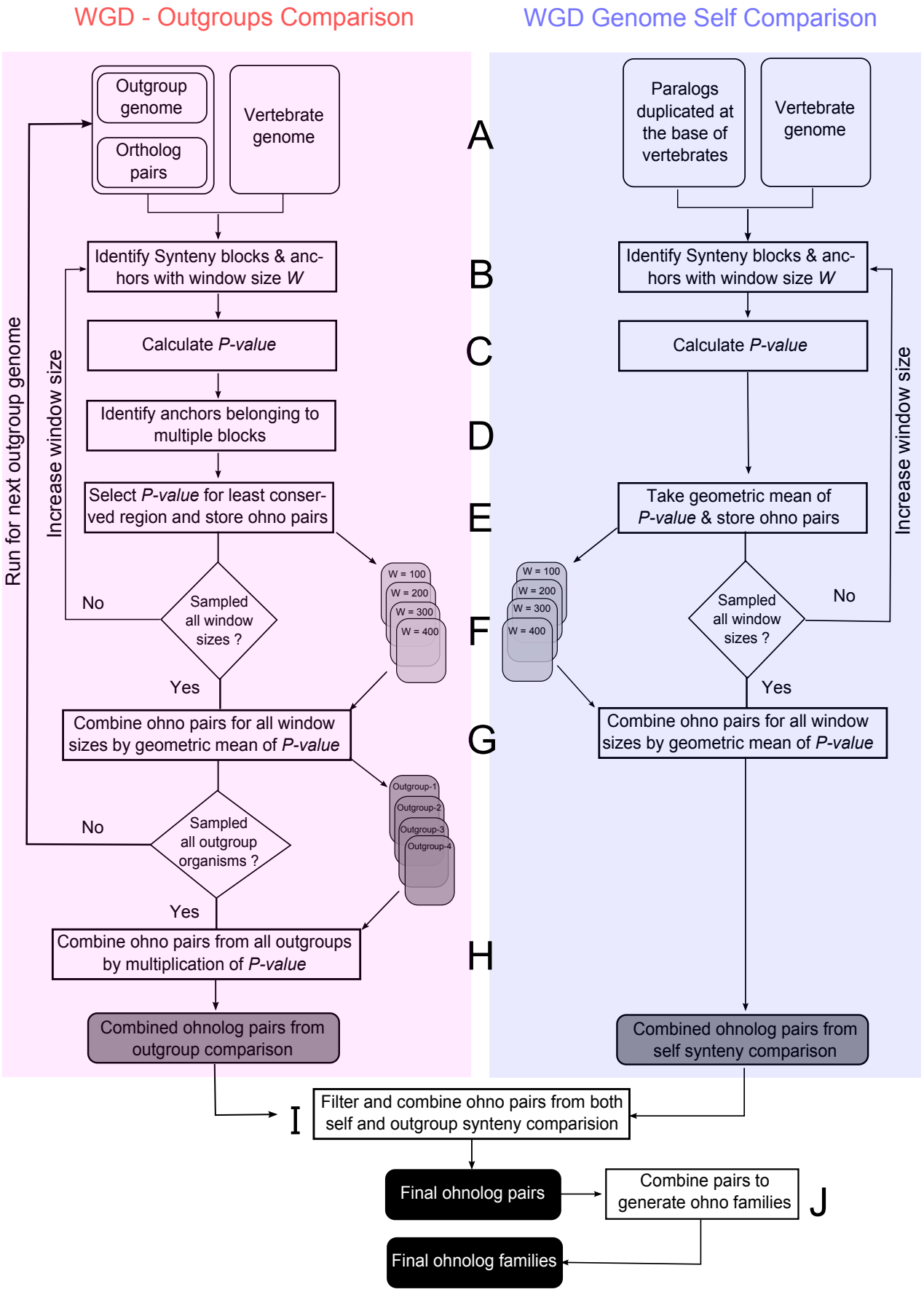

Figure 6.1: Flowchart of the algorithm to identify ohnolog pairs and construct ohnolog families for a single vertebrate genome by synteny comparison with multiple outgroup genomes (left panel) and self-comparison (right panel).

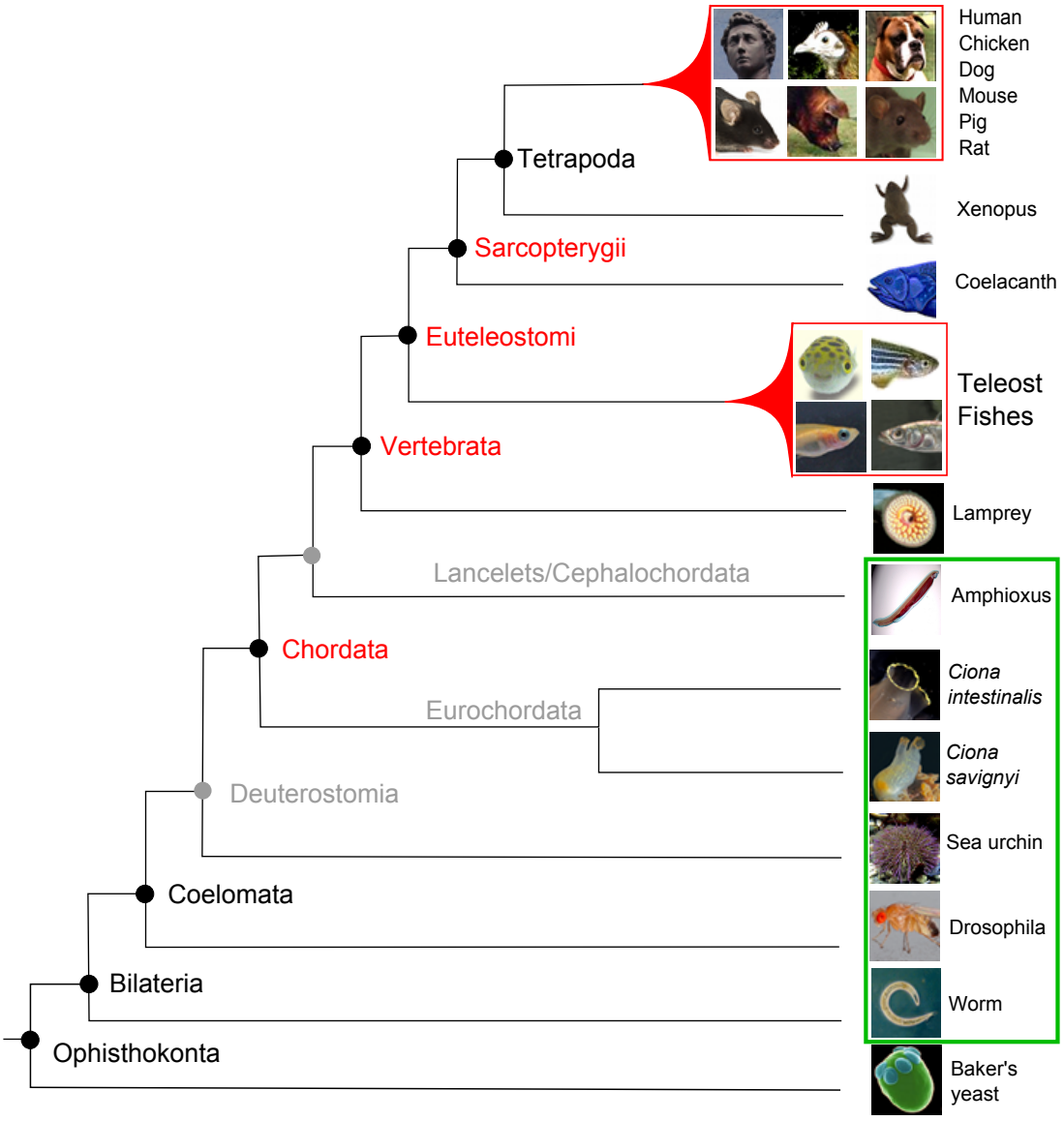

Figure 6.2: Schematic tree for the organisms analyzed in this study. The duplication nodes are from Ensembl Compara. Gray nodes are not the part of Ensembl. Paleopolyploid genomes are marked in red boxes and outgroups (for 2R-WGD) are in green box.

6.1.2 Orthologs and paralogs

We then identified vertebrate genes sharing the same outgroup ortholog using pairwise BLASTp (E-$value < 10^{-5}$) and selected the best outgroup match for each vertebrate gene. The vertebrate genes that share the same invertebrate ortholog are the only candidate gene duplicates considered in the vertebrate genomes.

Duplicated genes within vertebrate genomes (paralogs) and their relative node of duplication were obtained from Ensembl compara [Vilella et al., 2009] using BioMart. We noticed ambiguities in the relative duplication timing of genes in different Ensembl versions. This is because Ensembl compara assesses duplication times by constructing gene families through clustering, and then reconciles gene trees for each family with the species tree. Therefore, as new organisms are added in recent versions the duplication node of a gene can change. For example the paralog pairs *RalGDS – RGL4* has been annotated to Eutheria in Ensembl v66,

Bilateria in v67, again to Eutheria in v68-69 and to Euteleostomi in v70.

Therefore, rather than using just one node at the base of vertebrates, we considered paralogs from four duplication nodes at the base of vertebrates: Chordata, Vertebrata, Euteleostomi and Sarcopterygii, as candidate ohnologs if they also fulfil our synteny criteria. For the human genome, we further took the consensus of 6 Ensembl releases (v65 – v70) and collected paralogs whose relative duplication time were annotated to one of these four nodes in the majority of the recent releases, taking the most recent version (v70) in the case of ambiguities. Table 6.1 lists the number of genes, orthologs and paralogs for all the analyzed genomes.

Table 6.1: Number of protein coding genes, orthologs and paralogs for analyzed vertebrate (A) and invertebrate (B) genomes

A. Vertebrates									
Organism	Total genes	*Total paralogs	*Candidate paralogs	Ortholog pairs					
				B. floridae	*C. intestinalis*	*C. savignyi*	*S. purpuratus*	*D. melanogaster*	*C. elegans*
Homo sapiens	20415	127981	36775	16462	14746	13955	15807	14165	13135
Canis lupus familiaris	19574	103380	31402	16424	14731	13857	15756	14051	13011
Gallus gallus	15310	48334	16688	12006	10841	10201	11522	10359	9590
Mus musculus	22571	273095	48737	17707	15693	14795	16927	15053	13998
Rattus norvegicus	22865	218224	48678	18750	16635	15593	17998	16004	14832
Sus scrofa	19429	121315	29930	16414	14413	13544	15715	13744	12626
B. Invertebrate (Outgroup)									
Protein coding genes				50817	16658	11604	28525	13924	20505

* The numbers correspond to number of pairs

6.2 Identification of synteny blocks and anchors (6.1B)

A synteny block is defined as a region between an outgroup and a vertebrate genome (Figure 6.3 A), or two regions within the same vertebrate genome (Figure 6.3 B) having multiple homologous gene pairs. Between the genomes of two species such blocks represent conserved genomic regions descended from their last common ancestor. Within the genome of the same organism, synteny blocks represent duplicated sister regions, provided the duplication time of the genes residing on such blocks is the same. Vertebrate WGDs are among the oldest known genome duplications and the conservation of gene order or collinearity is limited [Putnam et al., 2008]. However, conservation of macro- or content-based synteny can be observed between genomic regions, where there is a statistical enrichment of orthologs, even after more than 500 million years of independent evolution since the divergence between vertebrates and invertebrates.

We used a window based approach to detect such regions between outgroup and vertebrate genomes extending earlier similar approaches [Dehal and Boore, 2005; Makino and McLysaght, 2010]. Any two regions between an outgroup and a vertebrate genome were considered to be candidate syntenic regions if there were at least m orthologous gene pairs between them, within a window of size W, where $2 \leq m \leq W$. We scanned the genomes of invertebrate and vertebrate organisms by placing a symmetric window around the ortholog genes in each genome in such a way that there are $^{W}/_{2}$ genes upstream and downstream (Figure 6.3 A). Hence, the ortholog partner under consideration was at the center of the windows in each genome. If there were at least 2 ortholog pairs in this window W, including the central pair, it was considered to be a synteny candidate (necessary but not sufficient condition). All such blocks were identified genome-wide and were labelled by the ortholog pair at the center of the blocks, referred to as the anchors (O_7–V_7, O_7–$V_7^{'}$, (Figure 6.3 A). At the chromosome boundaries, we kept the window size fixed by extending it in the opposite direction and making the

window asymmetric around the anchor gene to avoid biasing the calculation of synteny *P-value* as described in the next section.

The procedure was repeated between regions in the same genome to perform the self comparison of vertebrate genomes and to identify all *vertebrate-vertebrate* anchors (*e.g.* V_7–$V_7^{'}$, (Figure 6.3 B). While comparing two regions within the same vertebrate genome, we only considered paralogs duplicated at the base of vertebrates for each of the vertebrate genome according to Ensembl compara as detailed in Section 6.1.2.

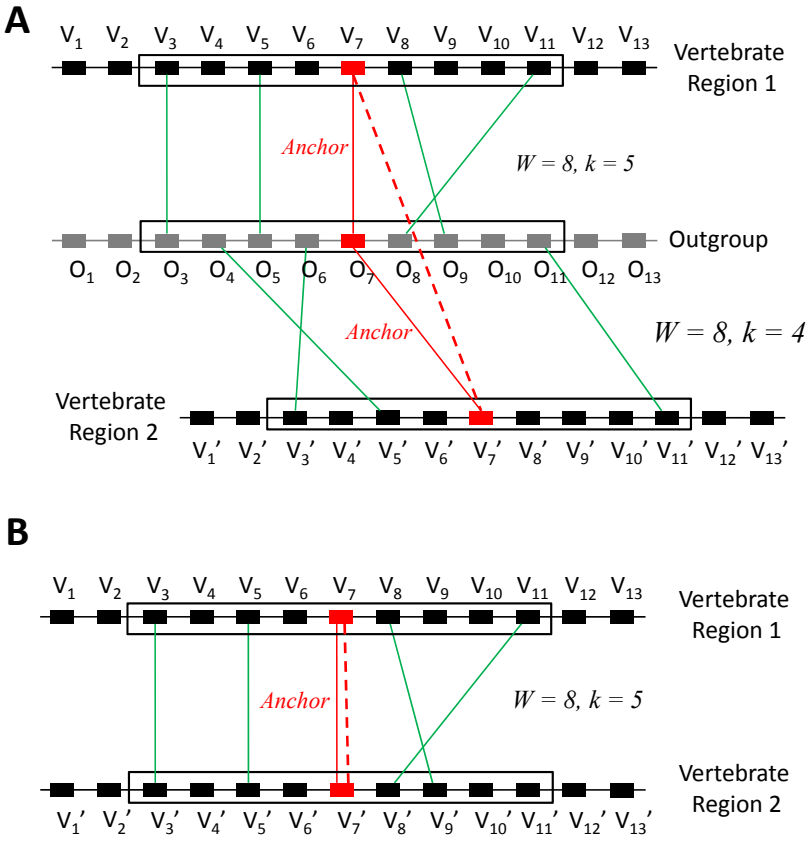

Figure 6.3: Comparison of genomic regions to identify anchor pairs (in red) and ohnolog candidate pairs (dashed red). Each block represents a gene labeled by O_i on the outgroup genome and V_i on the vertebrate genome. Duplicated regions in the vertebrate genome are marked by $V_1^{'} - V_n^{'}$. Other orthologous (A) and paralogous (B) relations are depicted by green lines.
(A) Identification of *anchors* between an outgroup window and two windows in the vertebrate genome. Using a window of size 8(+1) centered around O_7–V_7 and O_7–$V_7^{'}$ orthologous pairs, we observe 4 and 3 additional genes between the outgroup and vertebrate regions 1 and 2 respectively. Hence, O_7–V_7 and O_7–$V_7^{'}$ are the two *anchors*, and since both the vertebrate genes V_7 and $V_7^{'}$ share the same outgroup ortholog O_7, they are inferred as paralogs and make a candidate ohnolog pair.
(B) Identification of ohnologs between two regions in the same vertebrate genome. The anchor V_7–$V_7^{'}$ having four additional paralog pairs ($k = 5$) in the window can directly be taken as a candidate ohnolog pair.

6.3 Calculation of *P-value* to rule out spurious synteny (6.1C)

Since we resort to content based synteny and there is no bound on W and m, it is important to establish that the observed synteny is not just by chance especially for larger W and small m. Given the ortholog (or paralog) relations and location of the homologous gene on the outgroup and vertebrate genomes, we calculate the probability of finding at least k homologous genes by chance in a window ($P_{\geq k} \equiv P-value$) for all identified anchors, where k is the number of observed outgroup genes (excluding anchor) having orthologs in the vertebrate window under consideration.

For any gene i in a given synteny block (e.g. $i = O_5 : O_9$; Figure 6.4),we first calculate the probability P_i of finding at least one ortholog of gene O_i by-chance in a given window of size W_p in the vertebrate genome as follows

$$P_i = 1 - \frac{\sum_S (l_s - W_p + 1)}{N - W_p}$$

Where,
S = *Total segments of consecutive genes in the vertebrate genome without any ortholog of the outgroup gene* O_i *(e.g. blue segments of length* W_p *on the vertebrate genome between ortholog of* O_8 *in Figure 6.4)*
ls = *length of a stretch of consecutive genes in the vertebrate genome without any ortholog of* O_i
W_p = *window size in both genomes*
N = *total number of genes on the vertebrate chromosome under consideration*

Here $\sum_S (l_s - W_p + 1)$ represents the number of windows of size W_p in the vertebrate genome without any ortholog of an outgroup gene O_i (*e.g.* windows highlighted in blue in the schematic genome in Figure 6.4. We calculate P_i for all the genes in the outgroup window having orthologs in the vertebrate genome (Figure 6.4). These basic probabilities are used to calculate significance measures.

For every synteny block, we then calculate the probability of observing at least k orthologs by-chance between two given windows (around the anchor) in the vertebrate and outgroup genomes using combinations of P_i for all genes in the window ranging from 0 to $(k-1)$. For example, in Figure 6.4, there are three more ortholog pairs (green) between the two windows (boxes) in addition to the central anchor pair (red). Therefore, the *P-value* for the observed synteny is the probability that we find orthologs of any 3 or more genes in the same window of the vertebrate genome. We can use combinatorics to calculate the probability of finding *any* 3 or 4 genes in a given window by chance. For example, the probability of finding any 3 genes can be calculated using all combinations of the P_i of 3 genes. For realistic window sizes (100 to 500), however, the combinations exponentially increase the computational time and complexity. Therefore, we first assume that the P_i for each gene in the block is comparable, follow a mean-field approach and average the probability for all outgroup genes with orthologs in the vertebrate genome (P_{O5}, P_{O8}, P_{O9}) excluding the anchor (P_{O7}) which defines the window pairs centered around O_7 and V_{18}.

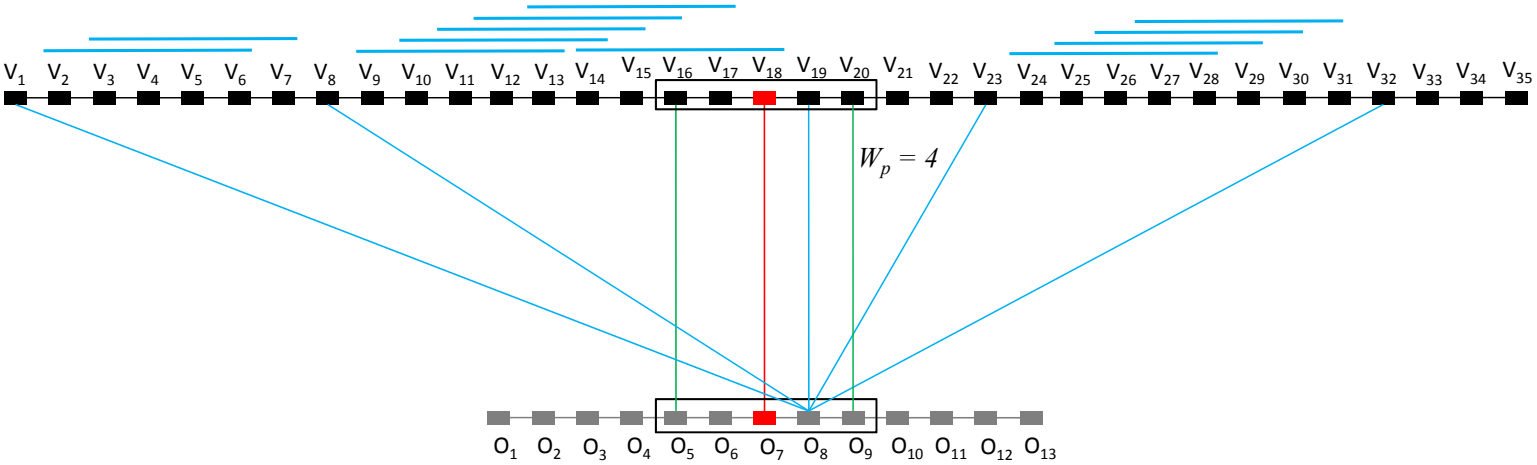

Figure 6.4: The calculation of P_i for an outgroup gene O_8. O_8 has 5 orthologs in the vertebrate genome: V_1, V_8, V_{19}, V_{23} and V_{32}. 12 windows without any of these orthologs are highlighted in blue for $W_p = 4$. P_i for this anchor then becomes 1 – 12/31 = 0.6, where 31 are the total possible windows on this schematic vertebrate genome ($N - W_p$).

$$log(\overline{P}) = \frac{1}{N_0} \sum_{i \neq anchor}^{N_0} log(P_i)$$

Where $\overline{P}$ is the geometric average of P_i (excluding the anchor) for the window under consideration. The probability of observing $\geq k$ genes where k is the observed orthologs between two windows can then be estimated using the following binomial formula.

$$P_{(\geq k)} = \sum_{j=k}^{N_0} \left[\binom{N_0}{j} \times \overline{P}^j \times (1-\overline{P})^{N_o - j} \right] = 1 - \sum_{j=0}^{k-1} \left[\binom{N_0}{j} \times \overline{P}^j \times (1-\overline{P})^{N_o - j} \right]$$

Where,
$P_{(\geq k)} \equiv P - value$ = probability of observing at least k homologs by chance between two windows
N_o = Number of outgroup genes in the outgroup window having at least one ortholog in the vertebrate genome (excluding the anchor pair)
k = observed number of orthologs between the two windows (excluding the anchor pair)

The same approach can be utilized while comparing two windows within the same vertebrate genome.

This *P-value* defines our confidence for each anchor pair between any two regions of the outgroup and vertebrate genomes. By definition, the *P-value* is affected by the number of orthologs (or paralogs) between the genomes being compared and the genomic context, *i.e.*, the distribution of orthologs of *anchor* and other genes residing in the same window.

6.4 Identify putative ohnolog pairs (6.1D)

Due to two rounds of genome duplication in the vertebrate genome, each synteny window in the outgroup genome should ideally correspond to up to four windows in the vertebrate genome,

however, only a few ohnologs are retained in > 2 copies. To identify such candidate ohnolog pairs, we identify anchors in the vertebrate genome that share the same outgroup gene (*e.g.* $O_7 - V_7$ and $O_7 - V_7^{'}$, Figure 6.3 A). Vertebrate genes belonging to such anchors thus become the ohnolog candidates. Yet, at this step the duplication time of ohnolog candidates can be incorrect. In fact, due to our very relaxed synteny criteria, many of these candidate ohnologs are not duplicated at the correct time in Ensembl. Therefore, we already exclude such putative ohnolog pairs that are not duplicated at the base of vertebrates, or if the pair does not exist in Ensembl. In later steps, we will further filter these ohnolog candidate pairs by combining their *P-values* (see Section 6.5).

For vertebrate genome self-comparison, since we have restricted ourself to paralogs at the base of vertebrates, anchors can be directly taken to be the candidate ohnologs if they pass our probability filters.

6.5 Combine *P-value* from anchors (6.1E)

Note that the *P-value* is calculated for each anchor pair. Therefore, for the candidate ohnolog pairs, we obtain two *P-values* corresponding to each anchor. For example, for the ohnolog candidate pair $V_7 - V_7^{'}$ in Figure 6.3 A, we obtain two *P-values* from anchors $P_{O_7-V_7}$ and $P_{O_7-V_7^{'}}$. Since the *P-value* depends on genomic context, these values can be very different. Therefore, for each ohnolog candidate pairs we take the *P-value* from the least conserved block to be the *P-value* of the ohnolog candidate pair. Hence,

$$P_{V_7-V_7^{'}} = Max.(P_{\geq k(O_7-V_7)},\ P_{\geq k(O_7-V_7^{'})})$$

Similarly, for self comparison, because self comparison is directional (each block being alternatively chosen to play the role of the outgroup region to calculate *P-vlaue* as described in 6.3) we also get two *P-values*, e.g. for $V_7 \rightarrow V_7^{'}$ and $V_7^{'} \rightarrow V_7$ comparison (Figure 6.3 B). We take geometric mean of these as the representative *P-value* for the ohnolog candidate pair $V_7 - V_7^{'}$.

$$log(P_{V_7-V_7^{'}}) = \frac{log(P_{\geq k(V_7 \rightarrow V_7^{'})}) + log(P_{\geq k(V_7^{'} \rightarrow V_7)})}{2}$$

6.6 Sample genomes with multiple window sizes (6.1F,G)

Typically in window based approaches for inferring synteny, only a single window is used [Dehal and Boore, 2005; Makino and McLysaght, 2010]. However, there is no optimum size of the window. Since we have a quantitative measure of the statistical significance of our comparison, we repeat the steps A-E in Figure 6.1 using multiple window sizes. We start with a relatively small window size of 100 and sample each genome with increasing window sizes of 200, 300, 400 and 500; with a minimum required number of orthologs/paralogs in each window, $k = 2$ (including the anchor).

If an ohnolog pairs is identified by multiple window sizes, we obtain a representative *P-value* for that pair using geometric mean of *P-values* from all the window sizes by which the pair is identified.

$$log\ \ (P_{(\geq k)\overline{W}}) = \frac{1}{n_W} \sum^{n_W} log(P_w)$$

Here, n_W is the total windows by which a pair can be identified, and $P_{(\geq k)\overline{W}}$ is the geometric average probability from all these windows. A pair can only be identified by larger windows or smaller windows (for pairs on the same chromosome). We compute the average *P-value* and store all the ohnolog pairs for a particular outgroup.

For self comparison, these pairs can now be filtered based on desired criteria.

6.7 Combine *P-value* from all outgroups (6.1H)

We perform comparison of each vertebrate genome with six different outgroups to overcome lineage specific rearrangements. If an ohnolog pair is identified by multiple outgroups, it strongly suggests that the pair is a 'true' ohnolog. Therefore, we combine ohnolog pairs from multiple outgroups using 'AND' rule by multiplying the *P-values* from all the outgroups. This amounts to assume that the synteny conservation is independent for different outgroups due to lineage specific rearrangements.

$$\textbf{\textit{P-value}} = \prod^{all\ outgroups} P_{(\geq k)\overline{W}_o}$$

This value hence, is the final *P-value* for a particular ohnolog candidate pair from all windows and outgroups. Using multiple outgroups thereby improves the statistical significance of the inferred ohnolog pairs.

6.8 Filter ohnolog pairs to remove false positives (6.1I)

For each pair that is identified by both the outgroup (at least one) and self comparison, we obtain two *P-values*. There are many pairs identified only by self or outgroup comparison. Any custom criteria can now be used to filter high confidence ohnologs. We filter ohnolog pair candidates **(1)** having duplication timing at the base of vertebrates, and **(2)** with three different synteny criteria combining *P-values* from both outgroup and self-comparison, as follows.

- **Strict criteria**: $P_{outgroup} < 0.01\ \ AND\ \ P_{self} < 0.01$
- **Intermediate criteria**: $P_{outgroup} < 0.05\ \ AND\ \ P_{self} < 0.3$
- **Relaxed criteria**: $(P_{outgroup} < 0.05)\ OR\ (P_{outgroup} < 0.5\ \ AND\ \ P_{self} < 0.01)$

These three sets represent decreasing confidence in the ohnolog status of the identified ohnolog pairs. In principle, the relaxed criteria may also include a number of paralogs from large scale segmental duplicates from the origin of vertebrates.

6.9 Construction of ohnolog families (6.1J)

Using the filtered ohnolog pairs from strict, intermediate and relaxed *P-value* criteria, we then built ohnolog families *i.e.* paralogous families of genes retained from vertebrate WGDs. Due to the two rounds of WGDs, we expect that most of these ohnolog families should consist in size 2, 3 and 4. However, SSDs and large scale segmental duplicates may lead to family sizes larger than four.

To construct ohnolog families, we start with an ohnolog pair and use depth first search [Tarjan, 1972] to traverse ohnolog pair space using both the genes as start nodes until no new ohnolog partner can be found for start nodes or for any of the branches[1]. Ohnolog families constructed using this exhaustive approach may contain genes which are SSD with respect to each other but are ohnolog partner for a third gene in the family.

We identify and merge such SSDs together. If any two genes within the same family reside on the same chromosome inside the smallest window in our comparison (*i.e.* within 100 genes) it is assumed to be an SSD. For all the genes within an ohnolog family, we also check if any of the possible pairs have duplicated earlier or later than the time of vertebrate WGD. We also merge such SSDs together.

We observed instances where two or more genes (or regions) are ohnologs with respect to the same third region, and yet duplicated later than the time of 2R-WGD. For example, in the human genome *TLX2* (on chromosome 2) and TLX3 (on chromosome 5) are both ohnologs with *TLX1* (on chromosome 10). In such instances, we count the number of unique pairs within the largest window between the two regions duplicated later and their ohnolog partner (e.g. *TLX1 - TLX2* and *TLX1 - TLX3*). We consider the window having more genes to be the original region that duplicated by WGD.

Therefore, the final duplication aware families consist of ohnolog partners along with the information on recent and/or old SSD if the ohnolog has undergone additional duplication episodes according to Ensembl family trees.

6.10 Randomization of the human genome

To verify that our algorithm to combine ohnolog information based on *P-value* incorporates a limited number of false positive ohnolog pairs, we perform the same comparison to identify ohnologs in randomized human genome. We shuffle the human genome to place each gene on a random chromosome at a random position. Keeping chromosome size, ortholog and paralog relations fixed, we repeat our approach to identify ohnologs and calculate combined *P-value* using randomized human and the original outgroup genomes for a moderate window size of 300. We then combine the *P-values* for both the cases by multiplication and compare the results of the randomized and original comparison (Results, section 9.1.1).

6.11 Small Scale Duplicates (SSD)

All the paralogs from Ensembl compara which we could not identify as ohnologs can be taken to be SSD duplicates. These SSD duplicates correspond to duplicates from all ages, before and after the 2R-WGD.

We also generated another dataset of SSDs based on sequence comparison. We ran an all-against-all BLAST*p* using entire human proteome, and selected the best non-self hit for each gene. If the best non-self hit does not correspond to one of the ohnolog partners of the gene, the pair can be taken to be an SSD pair.

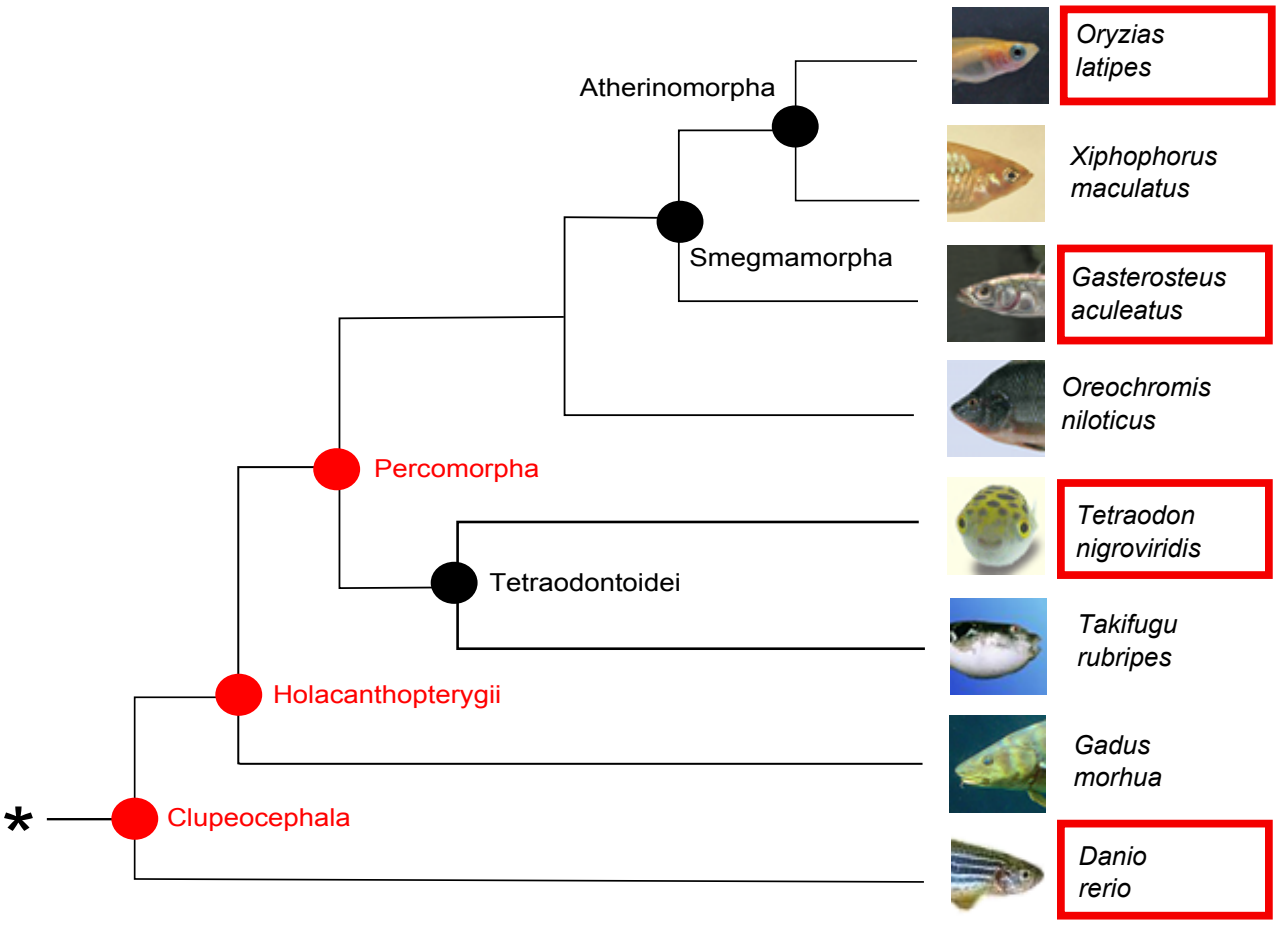

Figure 6.5: Schematic species tree of the sequenced fish genomes and the duplication nodes from Ensembl. Species marked by red boxes have assembled chromosomes and were analysed in this study. Paralogs from the three nodes marked in red were taken to be the candidates with correct duplication time.

6.12 Ohnologs in the teleost fish genomes: the 3R-WGD

Teleost fishes have experienced an additional whole genome duplication (3R-WGD) after divergence from tetrapods [Amores et al., 1998; Jaillon et al., 2004]. We also identified ohnologs from both the 2R and 3R-WGD in four sequenced teleost fish genomes where the genomes have been assembled into chromosomes: Zebrafish (*Danio rerio*), Tetraodon (*Tetraodon nigroviridis*), Stickleback (*Gasterosteus aculeatus*) and Medaka (*Oryzias latipes*).

6.12.1 The 2R-WGD

We used the same approach, as described for tetrapods in the previous sections to identify ohnologs from 2R-WGD in the four fish genomes. All the protein coding genes and their genomic coordinates for these four fish genomes were downloaded from Ensembl *v70*. The same six invertebrate outgroups were used and the fish orthologs sharing outgroup genes were also identified by running BLAST*p*. Paralogs for all the fishes were also obtained from Ensembl. However, paralogs from only three nodes: Chordata, Euteleostomi and Vertebrata were considered to have correct duplication timing for fish 2R-WGD.

[1] http://en.wikipedia.org/wiki/Depth-first_search

Synteny comparison between all outgroups and fish genomes and within the same fish genomes were performed using exactly the same approach. The three set of ohnologs and families were also constructed for all of the four fish genomes.

6.12.2 The 3R-WGD

Since the third WGD occurred around 350 MY ago, *i.e.* after the divergence of the tetrapods, the tetrapod genomes can be used as the outgroups to identify ohnologs retained from the third round of WGD. Recently, the genome of Coelacanth (*Latimeria chalumnae*) have been sequenced and added to Ensembl [Amemiya et al., 2013]. The Coelocanth genome is the only available fish genome in Ensembl that has diverged from the teleost fishes before the third round of WGD [Noonan et al., 2004; Amemiya et al., 2010]. Therefore, to identify ohnologs from 3R-WGD, we used human, chicken, dog, mouse, rat and coelacanth genomes as outgroups.

Teleost fish orthologs sharing the same tetrapod/coelacanth sequence were identified using BLAST*p*, and paralogs were downloaded from Ensembl. We used three Ensembl compara nodes near the base of the origin of teleost fishes: clupeoceptala, holocanthopterygii and percomorpha as the candidate 3R-WGD nodes Figure 6.5.

Self synteny comparison using paralogs belonging to these three nodes, and synteny comparison with the five tetrapods and coelacanth genomes were performed. Ohnologs from all the comparisons were filtered using the same three *P-value* criteria. Finally, we also constructed the ohnolog families from the genes retained from the 3R-WGD in all the four fish genomes.

6.13 Development of the OHNOLOGS server

To give easy access to these ohnolog pairs to the biologist community, we also developed a server named *OHNOLOGS* from where, as of now, all the ohnolog pairs and families from the 2R-WGD in the six tetrapod genomes can be retrieved.

The server was implemented in Perl-CGI and has been hosted on one of the virtual machines at Institut Curie. The server can be accessed at: `http://Ohnologs.Curie.fr`.

7

Collection of Cancer/Disease Genes & Functional Genomic Data

To assess the evolution of "dangerous" genes, we first collected a comprehensive set of human genes either known to be mutated in different diseases, or having certain functional properties, making them highly susceptible to deleterious mutations. This includes genes known to be mutated in different human cancers, genes implicated in genetic diseases other than cancer, genes having dominant negative or haplo-insufficient mutations in human diseases and genes with autoinhibitory folds. We also obtained information of other functional genomic properties known to affect the retention of genes after duplication e.g. essentiality, genes involved in formation of protein complexes, genes having variable copy numbers (CNV) in the human genome and expression level of genes in healthy human tissues. For the entire analysis Ensembl database (release-70) was used as a reference, and genes from all other databases were mapped to Ensembl.

7.1 Cancer genes

Candidate cancer genes were obtained from multiple databases. Table 7.1 lists details of these databases along with the methodology of obtaining the gene lists. Out of these 14 databases, Cancer Gene Census (CGC), Atlas of Over-expressed Genes In Cancer (AOGIC), UniProt Knowledge Base (UniProtKB: SwissProt) and Tumor Suppressor Gene database (TSGene) contain experimentally verified or manually curated cancer genes. These databases are also regularly updated. Other databases however, may include some false positives as the gene lists are either old, not regularly updated or parsed using text searches. Gene lists for AGCOH, TGFDB and TSGDB were downloaded from F-census [Gong et al., 2010] using their export facility, as they could not directly be downloaded from their respective websites.

Majority of cancer genes were obtained from the Catalogue Of Somatic Mutation In Cell (COSMIC) that contains genes found to be mutated in high-throughput cancer genome sequencing projects, along with the information on mutation type in different human primary tumor tissues.

Table 7.1: Cancer Databases & the Details of Obtaining Cancer Genes

Database Name	**Priority**	**Methods**	**Total genes**	**Onc/TS Status**	**Reference**
CGC	1	Downloaded and processed gene list from CGC homepage.	420	Yes	[Futreal et al., 2004]
AOGIC *v-1*	2	Downloaded and processed gene list from Supplementary information S2.	79	Yes	[Santarius et al., 2010]
UniProtKB: Swiss-Prot	3	Searched using: (keyword:"Oncogene [KW-0553]" OR keyword:"Proto-oncogene [KW-0656]") AND organism:"Human [9606]" and keyword:"Tumor suppressor [KW-0043]" AND organism:"Human [9606]"	360	Yes	[Magrane and Consortium, 2011]
TSGene	4	Downloaded gene list using download facility	637	Yes	[Zhao et al., 2013]
OMIM	5	Searched using keywords: "oncogene"[Title] OR "protooncogene"[Title] AND "has locus"[Properties] and "tumor suppressor"[Title] AND "has locus"[Properties]	1023	Yes	[Hamosh et al., 2005]
Ensembl (release 70)	6	Searched gene names using: tumor, cancer, oncogene/"proto oncogene" and "tumor suppressor" as keywords	332	Yes	[Flicek et al., 2013]
NCBI Gene	7	Searched using: "oncogene"[All Fields] OR "protooncogene"[All Fields] AND "homo sapiens"[ORGN] AND "current only"[Filter] AND "genetype protein coding"[Properties] and "tumor suppressor"[All Fields] AND "homo sapiens"[ORGN] AND "current only"[Filter]AND "genetype protein coding"[Properties]	1,236	Yes	[Maglott et al., 2011]
GeneCards *v-3*	8	Using advanced search: Aliases & Descriptions or Disorders or Gene Function or Summaries has "oncogene" or "tumor suppressor" AND Category: Protein Coding	775	Yes	[Safran et al., 2010]
TAG	9	All genes were downloaded and processed from website.	501	Yes	[Chen et al., 2013a]
TSGDB	10	Gene list was downloaded via F-census.	141	Yes	[Yang and Fu, 2003]
Candidate TS Genes	11	Candidate tumor suppressor genes were downloaded and processed from Table 1 in article.	154	Yes	[Volinia et al., 2008]
COSMIC *v-64*	12	Entire database was downloaded from COSMIC FTP and processed using custom perl scripts.	4,237	No	[Forbes et al., 2008]
TGFDB	13	Gene list was downloaded via F-census.	302	No	[Baxevanis, 2001]

Continued on next page...

Database Name	Priority	Methods	Total genes	Onc/TS Status	Reference
CancerGenes database	14	All genes were downloaded using export facility and processed using custom scripts.	3,058	No	[Higgins et al., 2006]

* Further information on these dataset can be found at http://www.oxfordjournals.org/nar/database/subcat/8/33. **CGC**: Cancer Gene Census, **AOGIC**: Amplified and Overexpressed Genes In Cancer, **TSGene**: Tumor Suppressor Gene Database, **OMIM**: Online Mendelian Inheritance in Man, **TAG**: Tumor Associated Gene Database, **TSGDB**: Tumor Suppressor Gene Database, **COSMIC**: Catalogue Of Somatic Mutations In Cancer, **F-Census**: A Database of Functional Census of Human Cancer Genes, **AGCOH**: Atlas of Genetics and Cytogenetics in Oncology and Haematology, **TGFDB**: The Tumor Gene Family of Databases

Many of these may simply be *passenger* mutations not affecting tumor progression significantly. Therefore, we downloaded entire COSMIC database and counted mutations in each gene across all mutated samples. Only genes having at least 15 non-synonymous mutations including at least one recurrent mutation were taken as candidate cancer genes from COSMIC.

Genes from all these databases were mapped to Ensembl *release-70* if the Ensembl id was not provided, using either Entrez gene id, gene name or symbol, or SwissProt id. The non-protein coding genes or genes that could not be matched to Ensembl were discarded. The number of genes from each database is listed in Table 7.1.

7.1.1 Oncogenes and tumor suppressors

We classified candidate cancer genes into oncogenes or tumor suppressors depending on mode of inheritance of mutations (dominant/recessive). Databases such as CGC, TSGene, AOGIC, TAG, TSGDB and SwissProt readily provide information on oncogenes and tumor-suppressor status of cancer genes. The same information could also be obtained from text searches and manual curation of related keywords in OMIM, NCBI gene and GeneCards databases. COSMIC database does not contain information about dominant or recessive nature of mutations. Therefore, we predicted this using a procedure adopted by [Bozic et al., 2010].

Oncogenes typically are activated in tumors by point mutations. Expression of tumor suppressor genes on the other hand, are silenced by degenerating mutations. Identification of activating point mutations is difficult, yet silencing mutations can be easily identified. Following Bozic et al., we counted type of mutations in all samples in COSMIC- *v64* using custom Perl scripts, and calculated ratio of inactivating mutations to other mutations. Inactivating mutations included *whole gene deletion, frame-shift insertion/deletion , nonsense substitution* and *complex frame-shift mutations*. If the ratio was greater than 0.2, the gene was considered to be a tumor suppressor, and oncogene otherwise. Using this approach most of the tumor suppressors from curated databases can correctly be classified [Bozic et al., 2010].

Different databases may disagree on oncogene/tumor-suppressor status of the same gene in many instances. Therefore to uniquely identify a gene as oncogene or tumor suppressor, each database was assigned a priority (Table 7.1) depending on level of manual and experimental curation or prediction from text searches and COSMIC. For each gene, the status was taken from the database with highest priority. We could finally classify 5,996 genes as oncogenes and 1,829 genes as tumor suppressors. If a gene was identified as both oncogene and tumor suppressor, we considered it as oncogene.

7.1.2 "Core" cancer genes

A high quality subset of all cancer genes containing genes which are either experimentally known to be implicated in cancer, or are highly mutated from tumor sequencing projects was then computed, called as "core" cancer genes. Core cancer genes consist of genes from CGC, SwissProt, TSGene and AOGIC, including COSMIC genes with at least 50 non-synonymous and one recurrent mutation. The purpose of this classification was to obtain a high confidence subset of cancer genes without any likely false positive candidates. A total of 2,743 genes were classified as core cancer genes with 1,932 oncogenes and 811 tumor suppressors.

7.2 Dominant & recessive disease genes

Genes implicated in Mendelian diseases and their inheritance patterns were also obtained from multiple resources. These genes primarily consist of genes mutated in genetic diseases, however, it may include some cancer genes having only somatic mutation. OMIM is the most comprehensive resource of human Mendelian disease genes. OMIM text file was downloaded on *27-June-2013* and processed using custom Perl scripts. Another set of human disease genes was obtained from GeneCards database [1]. For both OMIM and GeneCard genes, Ensembl ids were matched using MIM Ids and gene symbols respectively in Ensembl BioMart [2]. Disease genes were also obtained from two published reports: [Blekhman et al., 2008] & [Chen et al., 2013b]. Blekhman et al.'s study was based on a hand curated list of genes from OMIM (*h*OMIM). Chen et al.'s data includes *h*OMIM, other OMIM genes and genes from [Podder and Ghosh, 2011], and directly provided the Ensembl Ids. Ensembl Ids from Blekhman et al. were also obtained from BioMart using gene symbols. Reconciliation of all these datasets led to a total of 5,172 disease genes.

Inheritance pattern of these genes were obtained from OMIM and *h*OMIM. Perl regular expressions were used to parse *Inheritance* section from each OMIM entry. This section describes the inheritance pattern of diseases, which includes: dominant, recessive, multifactorial, polygenic, X/Y-linked, heterogeneous or unknown. The same gene can also be classified to be both dominant or recessive under different diseases or from different studies. We carefully curated all the inheritance patterns and obtained 679 and 888 genes unambiguously classified to be dominant and recessive disease genes respectively. Out of all genes, 446 genes had ambiguous inheritance patterns and 3,159 genes without the inheritance information.

7.3 Haploinsufficient and dominant negative genes

Haploinsufficient genes are the genes for which a single functional copy does not produce sufficient gene product, leading to a disorder. These genes are dosage sensitive. However, inactivating mutations have a dominant effect in these genes. Haploinsufficiency underlies many dominant diseases in humans including many cancers.

Dominant negative phenomena occurs when a mutated allele adversely interferes with the functional allele. The effect primarily occurs by dimerization between mutated and functional allele.

We obtained experimentally verified genes having dominant negative and haploinsufficient mutations from OMIM. Candidate genes were obtained from parsing OMIM text files with Perl

[1] http://www.genecards.org/cgi-bin/listdiseasecards.pl?type=full
[2] http://www.ensembl.org/biomart/martview/a16f093905d2e2f78a33931b6766eca0

regular expressions by searching for keywords mentioning these properties. Entire paragraphs having these keywords were put in a text file and then carefully reviewed to exclude any false positives or negative mentions. We could obtain 618 OMIM entries having mentions of haploinsufficiency and 901 entries including a reference to dominant negative mutations. Out of these, 525 and 670 OMIM entries were identified to be true positives.

BioMart was used to map the MIM accessions on Ensembl, yielding a total of 382 haploinsufficient and 566 dominant negative genes.

7.4 Genes with autoinhibitory protein folds

Most genes in the human genome code for proteins having multiple domains. In the encoded protein, domain-domian interactions are crucial for proper functions. Autoinhibitory proteins are also multidomain proteins where a part of protein interacts with the domain having the functional property. Such interaction keeps a protein in inactive form, unless this autoinhibition is counteracted by conformational changes triggered by events such as interactions with other partners, proteolysis or phosphorylation (Figure 7.1) [Pufall and Graves, 2002] . Inside cells, autoinhibited proteins are tightly regulated in signalling cascades, however, mutations may disrupt this inhibition leading to permanent activation of these proteins and diseases. Autoinhibition therefore, is a very important and widespread regulatory mechanism in cells. A few examples of autoinhibitory proteins include most tyrosine kinases [Hubbard et al., 1998; Hubbard, 2002, 2004], phosphatases [Denu and Dixon, 1998], Guanine Exchange factor (GEF) and GTPase Activating Proteins (GAP) from many growth hormone signalling pathways including RAS-RAL signalling pathway [Bodemann and White, 2008].

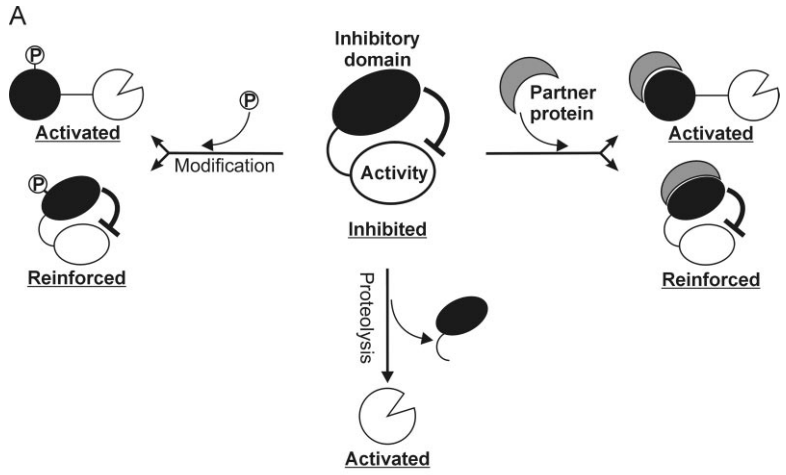

Figure 7.1: **Mechanism of Autoinhibition**. An autoinhibitory protein in its inhibited confirmation (centre). The autoinhibition can be activated and reinforced by mechanisms such as phosphorylation (left), interacting with a partner protein (right) or proteolysis (bottom). Adapted from [Pufall and Graves, 2002]

.

Although there has been hundreds of experimental studies identifying autoinhibitory domains in many human proteins, there is no systematic resource to obtain this information. Therefore, to obtain genes coding for proteins with autoinhibitory folds, PubMed was searched with keyword “autoinhibitory domain” on *19-November-2010* producing a total of 502 references having this keyword. Abstracts of these results were then carefully reviewed to identify “true” autoinhibitory genes. In cases where the information was not clear in the abstract, en-

tire paper was read to identify any verified autoinhibitory proteins or domains. Using protein or gene names mentioned in the text, we obtained Ensembl Ids manually. A total of 373 genes could be identified to be autoinhibitory using this approach.

Further gene candidates with autoinhibitory folds were obtained from databases, NCBI Gene, OMIM, SwissProt and GeneCards. Search terms: 'auto/self-inhibit*' was used on their respective homepage to identify candidate genes. Careful manual curation of these lists yielded 81/106, 50/60, 90/109, 167/273 true positives from NCBI Gene, OMIM, SwissProt and GeneCards respectively. Reconciling data from all these sources led to a total of 461 genes with autoinhibitory folds.

7.5 Genes coding for protein complexes

To test the impact of dosage balance on ohnolog retention we also collected genes involved in formation of verified protein complexes in the human genome. Relative expression of genes or dosage of interacting proteins is considered to be important for proper functioning [Liang and Fernandez, 2008]. Subunits of protein complexes in particular, are argued to be under strong dosage constraints as relative imbalance of the level of proteins may lead to malformed complexes detrimental to fitness [Birchler and Veitia, 2010; Veitia, 2009].

We obtained experimentally verified and manually curated protein complexes from Human Protein Reference Database **(HPRD)** [Keshava Prasad et al., 2009], COmprehensive ResoUrce of Mammalian protein complexes **(CORUM)** [Ruepp et al., 2010], **Census of Soluble Human Protein Complexes** [Havugimana et al., 2012] and **Gene Ontology** [Ashburner et al., 2000]. For all these databases, Ensembl Ids were matched after careful manual curation as described in sections below. Complex partners without Ensembl mapping were discarded. Our objective was to obtain a high confidence set of genes involved in formation of protein complexes, and therefore we did not include high throughput studies on protein-interactions in our analysis. Dosage balanced genes were defined as the genome coding for protein complexes or haploinsufficient genes Section 7.3.

7.5.1 Human protein reference database

HPRD [3] is a resource of manually curated information on human protein complexes from published literature. Total 1,521 protein complexes were downloaded from HPRD *Release-9*, which is based on RefSeq Id from NCBI. BioMart was used to obtain Ensembl-to-RefSeq mapping for their constituent genes (2,727 genes).

7.5.2 COmprehensive ResoUrce of Mammalian protein complexes

CORUM [4] is another manually curated repository of experimentally characterized protein complexes for mammals. Ensembl Ids for 5,947 CORUM complexes were obtained using NCBI Gene Ids provided in the downloaded file using custom Perl scripts (2,500 genes).

[3] http://www.hprd.org/index_html
[4] http://mips.helmholtz-muenchen.de/genre/proj/corum

7.5.3 Gene ontology

Gene Ontology [5] also contains expert curated biomedical ontologies. Term **GO:0043234** for protein complex ontology was searched in AmiGo browser [Carbon et al., 2009], to obtain all gene products associated to protein complexes according to Gene Ontology (total of 4,198 GO terms). BioMart was used to associate Ensembl gene ids to these gene ontology ids leading to 2,432 protein complex genes.

7.5.4 Census of soluble human protein complexes

Finally, we downloaded [6] 602 soluble protein complexes having 3,006 soluble proteins from [Havugimana et al., 2012]. Ensembl Ids for these complexes were obtained from SwissProt Ids provided in the downloaded file using BioMart yielding 2,996 genes.

7.5.5 Permanent complexes

As no distinction between transient and permanent complexes is made in all the above databases, we obtained a manually curated data set of permanent complexes from [Zanivan et al., 2007]. 11 permanent and 21 transient complexes were obtained from *Additional data file 10* and Ensembl Ids for 239 permanent and 308 transient complex genes were obtained from NCBI Gene Ids. A detail of these complexes is listed in Table 7.2.

Gene products participating in permanent complexes are only functional within their complexes, associated with obligatory complexes e.g. ATP_F0 and ribosomal subunits. However many cellular assemblies such as centrosome and replication complexes are transient and are easy to be disassociated. Our hypothesis was that genes in permanent complexes must show even higher susceptibility to relative dosage as compared to transient complexes.

Genes were considered to be a “complex” genes if they were included in any one of the five datasets (total 6,119 complex genes).

7.6 Essential genes

Essential genes are the genes which are critical for organismal/cellular health and on silencing mutations, lead to lethality or sterility. Essentiality has also been known to affect ohnolog retention [Makino et al., 2009]. Since no direct information of human essential genes is available, two different approaches were used to collect genes associated with essentiality in the human genome.

7.6.1 Human orthologs of Mouse essential genes

We used phenotype section of Mouse Genome Informatics[7] (MGI) database *version MGI_4.42* [Eppig et al., 2012]. MGI is a comprehensive resource of mouse functional genomics and gathers data from published literature and screens on phenotypic effect of different mutations. We searched mouse mutant alleles matching mammalian phenotype ontology terms: prenatal lethality (MP:0002080), perinatal lethality (MP:0002081), post natal lethality (MP:0002082),

[5] http://www.geneontology.org/GO.doc.cc-prot-complex.shtml
[6] http://human.med.utoronto.ca
[7] http://www.informatics.jax.org/phenotypes.shtml

Table 7.2: Details of 32 permanent & Transient Complexes from [Zanivan et al., 2007]

Symbol	Complex Type	Full Name	Total Subunits
ATP_F0	Permanent	ATP synthase, H+ transporting, mitochondrial F0 complex	11
ATP_F1	Permanent	ATP synthase, H+ transporting, mitochondrial F1 complex	5
COX	Permanent	Cytochrome c oxidase	11
SRS	Permanent	Small ribosomal subunit	32
LRS	Permanent	Large ribosomal subunit	43
MLRS	Permanent	Mitochondrial large ribosomal subunit	48
MSRS	Permanent	Mitochondrial small ribosomal subunit	30
Proteasome	Permanent	Proteasome Complex	31
PD	Permanent	Pyruvate dehydrogenase	8
RNA Pol II	Permanent	RNA polymerase II	12
RNA Pol III	Permanent	RNA polymerase III	9
AP2	Transient	Adaptor-related protein complex 2	5
APC	Transient	Anaphase promoting complex	9
Arp2-3	Transient	Arp2-3 Complex	7
ARC	Transient	Axin related complex	5
Centrosome	Transient	Centrosome Complex	64
Dynactin	Transient	Dynactin Complex	11
Exocyst	Transient	Exocyst Complex	9
Exosome	Transient	Exosome Complex	8
FA	Transient	Focal adhesion	54
GTC	Transient	Golgi transport complex	8
Nucleopore	Transient	Nucleopore complex	33
Nucleosome	Transient	Nucleosome Complex	32
ORC	Transient	Origin recognition complex	6
RFC	Transient	Replication complex	5
SRP	Transient	Signal recognition particle	6
SCF	Transient	Skp1-cull-F-box complex	3
SNARE	Transient	SNARE Complex	7
SWI/SNF	Transient	SWItch/Sucrose Non Fermentable (Nucleosome Remodeling)	13
TAFIID	Transient	Transcription Factor IID Complex	17
TRAPP	Transient	Trafficking protein particle complex	7
VHL	Transient	Von Hippel-Lindau complex	5

premature death (MP:0002083) and infertility (MP:0001924), generated by either knock-out, random gene disruption or gene trap mutagenesis [Liao and Zhang, 2008]. These genes represent a set of mouse essential genes. Alleles in which the above loss-of-function mutations resulted in other phenotypes (neither lethal nor infertile) were considered to be non-essential genes. Ensembl Ids for these essential and non-essential mouse genes were obtained using report file provided by MGI.

Human *1-to-1* orthologs (to remove effects of gene duplication in one of the lineages) of these mouse genes were obtained using BioMart. Discarding genes without a clear *1-to-1* mouse orthologs we obtained 2,938 human genes having an essential mouse ortholog and 3,498 genes having non-essential mouse orthologs (6,436 genes tested for essentiality in mouse).

7.6.2 Human essential genes from *In-vitro* knock-out experiments

Another set of human essential genes were obtained from Online Gene Essentiality Database **(OGEE)** [8] [Chen et al., 2012]. OGEE collected human essential genes from [Silva et al., 2008], who performed knock-out experiments using RNAi screening in five human mammary cell lines. Ensembl Ids obtained from OGEE were mapped to version-70 using BioMart, leading to

[8] http://ogeedb.embl.de/

a total of 18,938 protein coding genes tested for essentiality, with 2,701 essential and 16,237 non-essential human genes.

7.7 Genes with copy number variations

Copy number variations (CNV) are structural variants in the individual human genomes having different copies of a part of DNA. They represent a form of small scale duplications not yet fixed in the population. To study the effect of CNV genes on ohnolog retention [Makino and McLysaght, 2010], genes having CNV regions were obtained from Database of Genomic Variants **(DGV)** [Zhang et al., 2006] [9]. DGV collects copy number gains (CGV) or losses (CLV) data from published studies and reports the genomic coordinates of the variants. The variant file was downloaded from DGV, and using genome start-end coordinates of genes from Ensembl v-70, we identified human genes whose entire coding sequence fell within one of the CNV regions (either CLV or CGV). A total of 5,185 CNV genes were identified using this approach.

7.8 Expression Level

Absolute expression level of a gene has also been reported to be a determinant of gene retention after WGD [Seoighe and Wolfe, 1999; Gout et al., 2009, 2010]. Therefore, expression level of human genes in 78 healthy tissues and cell types were obtained from BioGPS[10] [Wu et al., 2009], which provides GC content adjusted-robust multi-array (GC-RMA) normalized expression levels (*Average Difference or AD values*) from microarray experiments [Su et al., 2004]. The expression data was based on *Affi*-U133-Plus2 microarray from Affimetrix.

Affymatrix tags and the expression values for all tissues were downloaded in form of a CSV file from BioGPS. These tags are small stretches of gene sequence used as reporters of expression levels for the corresponding gene. Mapping from these tags to NCBI Gene Id, UniGene Id or gene symbol were obtained by annotation file provided by BioGPS. Another set of direct mapping between Ensembl and *Affi*-U133-Plus2 tags was obtained using BioMart. Annotation file from Gene Expression Omnibus (GEO) was also referred to map tags to NCBI Gene Id. Using the three resources, a total of 31,154 out of 44,775 tags could be mapped to at least one of the gene Id/symbol. For tags mapping to other datbase Ids, corresponding Ensembl Ids were then obtained using BioMart resulting in 28,2237 tags matching an Ensembl gene.

These tags are regularly updated and to remove tags which could map to multiple genes or had become obsolete, we only kept tags having "..._at" suffix, which is used to characterize best quality tags. For each gene having multiple associated tags, the expression value from all mapping tags was averaged for each tissue. Finally, we used the median value of expression for all of the 78 healthy tissues as the representative expression level for each gene. We could obtain expression levels for 13,425 of a total 20,415 protein coding genes using this approach.

7.9 Disease genes in other vertebrates

We also obtain the genes implicated in cancer and diseases for the two mammalian model organisms as follows.

[9] http://dgv.tcag.ca/dgv/app/faq?ref=NCBI36/hg18
[10] http://biogps.org/#goto=welcome

7.9.1 Mouse

Mouse genes associated with OMIM diseases were downloaded from Mouse Genome Informatics (Report: MGI_OMIM.rpt)[11]. Mouse Ensembl ids were obtained using MGI Id to Ensembl id mapping from BioMart leading to a total of 1,242 mouse disease genes.

7.9.2 Rat

Genes implicated in cancer and other genetic diseases in the Rat genome were obtained from Rat Genome database (RGD) using RatMine [Laulederkind et al., 2013] [12]. A total of 1,921 genes implicated in diseases such as respiratory, neurological and cardiovascular disorders and cancer could be mapped to Ensembl ids for the rate genome using gene symbols from BioMart.

Table 7.3: Summary of gene counts in different categories

CAREGORY	TOTAL GENES		
All protein coding	20415		
All cancer	8899	Oncogene	5996
		Tumor-Suppressor	1829
Core cancer	2743	Oncogene	1932
		Tumor-Suppressor	811
All disease	5172	Dominant	679
		Recessive	888
Haploinsufficient	382		
Dominant negative	566		
Autoinhibitory	461		
Protein complexes	6119		
Permanent complexes	239		
Expression level	13425		
Essential genes (mouse orthologs)	6436	Essential	2938
		Non-essential	3498
Essentiality genes (*in-vitro* experiments)	18938	Essential	2701
		Non-essential	16237
Mouse disease genes	1242		
Rat disease genes	1921		

7.10 Analysis of ohnologs conservation using Ka/Ks ratios

To quantify the susceptibility of ohnologs *versus* nonohnologs to deleterious mutations through comparative sequence analysis, we used Ka/Ks ratio estimates (also called D_N/D_S) [Yang and Nielsen, 2000], which measure the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks). Ka/Ks ratios were calculated to evaluate the selection pressure on human ohnolog and nonohnolog genes and their respective invertebrate orthologs as follows. Protein sequences for *B. floridae*, *C. intestinalis*, *C. savignyi*, *D.*

[11] ftp://ftp.informatics.jax.org/pub/reports/index.html
[12] http://ratmine.mcw.edu/ratmine/bag.do

melanogaster, *C. elegans*, *S. purpuratus* and *H. sapiens* were obtained as described in Section 6.1.1 and [Singh et al., 2012]. These genes were based on Ensembl *release 61* [Singh et al., 2012]. Human-invertebrate orthologs were identified using BLAST*p* ($E-value :< 10^{-7}$). The best human-to-invertebrate hits were identified and used for Ka/Ks ratio calculations [Yang and Nielsen, 2000]. Each human ohnolog pair and human-invertebrate ortholog pair was aligned using clustalW [Larkin et al., 2007]. Ka and Ks values (Yang and Nielsen, 2000) have been calculated using the KaKs_Calculator 2.0 [Wang et al., 2010a]. Ohnolog or ortholog pairs for which Ka/Ks ratio could not be calculated (0.1%–0.7% of pairs) or having saturated Ks values (2%–7% of pairs) were discarded from the analysis. As these calculations were performed on the older release of Ensembl, these ratios were mapped to *release 70*.

8 Causal Inference Analysis

MULTIPLE genomic and functional properties are known to affect the retention of genes after duplication. In particular, dosage balance constraints have been argued to underlie the observed antagonistic retention pattern of genes in different functional categories after WGD and SSD. In this work, we provide evidence (in Chapter 10) that the human disease-related genes have also retained most of their duplicates from WGD but not from SSD. Therefore, susceptibility to dominant deleterious mutations also influences retention of ohnologs. To assess the relative contribution of each of these properties in the context of the retention of ohnologs, we needed a framework where the relative importance of each of these properties could be evaluated quantitatively. Rather than study two property correlation, we made the effort to disentangle the *indirect* effects from the *direct* effects of these functional constraints on the resulting biases in ohnolog retention.

8.1 Mediation Analysis

To this end, we have performed a Mediation analysis following the approach of Judea Pearl [Pearl, 2001, 2009, 2012a,b]. The Mediation framework, developed in the context of causal inference analysis [Pearl, 2009], aims at uncovering, beyond statistical correlations, *causal* pathways along which *changes* in multivariate properties are transmitted from a cause, X, to an effect, Y. More specifically, a Mediation analysis assesses the importance of a mediator, M, in transmitting the indirect effect of X on the response $Y \equiv Y(x, m(x))$, as shown in the Mediation diagram (Figure 8.1).

8.1.1 Total, direct & indirect effects

In mediation the framework, the total effect denotes the strength of the effect a causal variable X has on an outcome Y. However, a significant fraction of the total effect (TE) may not be direct, instead can be indirectly mediated by a third variable called a Mediator (M). Therefore, the term "direct effect" (DE) in a Mediation model quantifies the effect that is not affected by any other variable (mediators) in the model [Pearl, 2001].

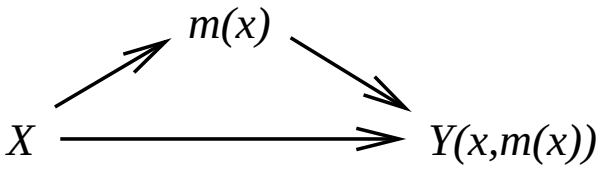

Figure 8.1: The Mediation Diagram

The direct and M-mediated effects of a control variable X to a change $x \rightarrow x'$, on an outcome Y can be quantified using counterfactual expressions, which formally enable to decouple the direct (x_1) and indirect (x_2) conditions on X seen by the outcome Y, *i.e.*, $Y(x_1, m(x_2))$. Hence, the direct effect of X on Y, $DE_{xx'}$, can be defined as the hypothetical change that Y would experience, if x could be changed to x' while keeping the mediator M to its original value $m(x)$. Likewise, the indirect effect of X on Y, $IE_{xx'}$, corresponds to the hypothetical change that Y would experience, if the mediator M could be changed to $m(x')$, while keeping the direct influence of X on Y to its original value x. This yields the following counterfactual definitions [Pearl, 2001] for the direct ($DE_{xx'}$), indirect ($IE_{xx'}$) and total ($TE_{xx'}$) effects on Y to a change $x \rightarrow x'$,

$$\begin{aligned} DE_{xx'} &= Y(x', m(x)) - Y(x, m(x)) \\ IE_{xx'} &= Y(x, m(x')) - Y(x, m(x)) \\ TE_{xx'} &= Y(x', m(x')) - Y(x, m(x)) \end{aligned}$$

where $DE_{xx'}$, $IE_{xx'}$ and $TE_{xx'}$ can be evaluated within the framework of Bayesian statistics [Pearl, 2001, 2009], using the counterfactual expressions,
$Y(x_1, m(x_2)) = \sum_m E(Y|x_1, m)P(M|x_2)$,

$$\begin{aligned} DE_{xx'} &= \sum_m \Big[E(Y|x', m) - E(Y|x, m)\Big] P(M|x) \\ IE_{xx'} &= \sum_m E(Y|x, m)\Big[P(M|x') - P(M|x)\Big] \\ TE_{xx'} &= E(Y|x') - E(Y|x) \end{aligned}$$

Note, in particular, that the total effect, $TE_{xx'}$, is not, in general, the simple sum of direct and indirect effects. Indeed, $TE_{xx'} = DE_{xx'} - Y(x', m(x)) + Y(x', m(x') \neq DE_{xx'} + IE_{xx'}$ for non-linear systems. Instead, the total effect $TE_{xx'}$ is related to the indirect effect of the *reversed* transition changing x' into x, *i.e.* $IE_{x'x} = Y(x', m(x)) - Y(x', m(x'))$, that is,

$$TE_{xx'} = DE_{xx'} - IE_{x'x}$$

8.1.2 Mediation calculations

We have applied the Mediation analysis to genomic data using a recent approach developed by Judea Pearl for non-linear systems using binary categories [Pearl, 2012a,b] (X, M, Y=0/1; $DE, IE, TE \in [-1, 1]$). To discriminate between direct and indirect effects in the case of binary

categories, Pearl's Mediation Formulae for direct effect (DE), indirect effect (IE) and total effect (TE) simply read, (page 23 in [Pearl, 2012a]),

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{8.1}$$
$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{8.2}$$
$$TE = f_1 - f_0 = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{8.3}$$

where f_0, f_1, g_{00}, g_{01}, g_{10}, g_{11}, h_0 and h_1 are obtained from the association table of the binary variables X, M and Y.

Table 8.1: Association table of binary variables to calculate direct and indirect effects

X	M	Y	n_{xmy}	$f_x = E(Y\|x)$	$g_{xm} = E(Y\|x,m)$	$h_x = P(M\|x)$
0	0	0	n_1	$f_0 = \frac{n_2+n_4}{n_1+n_2+n_3+n_4}$	$g_{00} = \frac{n_2}{n_1+n_2}$	$h_0 = \frac{n_3+n_4}{n_1+n_2+n_3+n_4}$
0	0	1	n_2			
0	1	0	n_3		$g_{01} = \frac{n_4}{n_3+n_4}$	
0	1	1	n_4			
1	0	0	n_5	$f_1 = \frac{n_6+n_8}{n_5+n_6+n_7+n_8}$	$g_{10} = \frac{n_6}{n_5+n_6}$	$h_1 = \frac{n_7+n_8}{n_5+n_6+n_7+n_8}$
1	0	1	n_6			
1	1	0	n_7		$g_{11} = \frac{n_8}{n_7+n_8}$	
1	1	1	n_8			
n_5+n_6 $+n_7+n_8$	n_3+n_4 $+n_7+n_8$	n_2+n_4 $+n_6+n_8$	$\sum_1^8 n_i$	$f = \frac{n_2+n_4+n_6+n_8}{\sum_1^8 n_i}$		$h = \frac{n_3+n_4+n_7+n_8}{\sum_1^8 n_i}$

8.1.3 Interpretation of Mediation results

The validity of the assumed Mediation model can be quantified in terms of the total effect of X on the outcome Y. A large total effect implies the strong overall causal effect of X on Y. Although $DE + IE \neq TE$, owing to possible non-linear couplings between direct and indirect causal effects [Pearl, 2001], the direct and indirect effects can be interpreted in terms of proportions of the total effect, DE/TE and IE/TE, as well as necessary and sufficient contributions from the direct and indirect causal pathways [Pearl, 2012a,b]. Namely, for a total effect TE of X on Y, the Mediation Analysis assesses that,

- the direct effect $X \to Y$ is *sufficient* as sole cause to account for a proportion Dir. $= DE/TE$ of the total effect, while,
- the indirect effect $X \to M \to Y$ is *necessary* as complementary cause to account for a proportion $1 - DE/TE$,

and, likewise,

- the indirect effect $X \to M \to Y$ is *sufficient* as sole cause to account for a proportion $Ind. = IE/TE$, while,
- the direct effect $X \to Y$ is *necessary* as complementary cause to account for the proportion $1 - IE/TE$ of the total effect.

Hence, the Mediation analysis [Pearl, 2001, 2009, 2012a,b] yields the following summary table for direct and indirect effects and their interpretations in terms of sufficient or necessary causes,

Table 8.2: Interpretation of Direct & Indirect Effects

$\nearrow M \searrow$ $X \longrightarrow Y$	direct effect $X \longrightarrow Y$	indirect effect $X \rightarrow M \rightarrow Y$	non-linear combination of direct and indirect effects
sufficient (as sole cause)	DE/TE	IE/TE	if $DE + IE \neq TE$
necessary (as complementary cause)	$1-IE/TE$	$1-DE/TE$	

8.1.4 Application on genomic properties

We applied the Mediation formula to assess the relative importance in terms of the total, direct and the indirect effects of various genomic properties on the retention of ohnologs. A numerical example to assess the direct effect of susceptibility to cancer mutations, $cancer = X$, and its mediation by dosage balance, $dosage.bal. = Z$, on the retention of ohnologs, $ohno = Y$ is described below. Dosage balanced genes contain 6,119 genes implicated in protein complexes, and cancer genes consist of 8,899 genes from all cancer gene set. 7,351 Ohnologs are from the relaxed *P-value* criteria.

First, we interpret each property as a binary variable e.g. $Y = 1$ reads that the gene is ohnolog and $Y = 0$ that it is not an ohnolog. Similarly, $X = 1$ implies that the gene is susceptible to cancer mutations; and $X = 0$, that it is not a cancer gene; and likewise for the mediator Z.

Then we compute the association table of the above three properties to keep treatment or mediator fixed at 0 or 1, and assess the retention of ohnologs for each of these combinations, as depicted below.

$dosage.bal$	$cancer$	$Ohno$	n_{xzy}	f_x	g_{xz}	h_x
0	0	0	6567	33.6% Ohno	25.7% Ohno	**38.2% Cancer**
0	0	1	2271			
0	**1**	0	2919		46.5% Ohno	
0	**1**	1	2539			
1	0	0	1834	41.5% Ohno	31.5% Ohno	**56.2% Cancer**
1	0	1	844			
1	**1**	0	1744		49.3% Ohno	
1	**1**	1	1697			
6119	8899	7351	20415	36.0% Ohno		**43.6% Cancer**

Hence, from the above association table we find that $f_0 = 0.336$, $f_1 = 0.415$, $g_{00} = 0.256$, $g_{01} = 0.465$, $g_{10} = 0.315$, $g_{11} = 0.493$, $h_0 = 0.381$ and $h_1 = 0.562$. Substituting this data in equation 8.1, 8.2 and 8.3 yields $TE = 0.0788$, $DE = 0.0467$ and $IE = 0.376$, and the fractions $DE/TE = 59.2\%$ and $IE/TE = 47.7\%$. This example, along with other mediation models has been interpreted in details in Chapter 12.

Part III

Results

9 Characterization of Vertebrate Ohnologs

OHNOLOGS were identified in six vertebrate (Tetrapod) genomes (Human, Chicken, Dog, Pig, Rat and Mouse) using the *P-value* method to detect ohnologs (Chapter 6). The *OHNOLOGS* program was run for all six vertebrates using six different invertebrate outgroups (Section 6.1). Each genome was scanned with five window sizes ranging from 100 to 500. Similarly, comparison of each vertebrate genome with itself was performed using the same approach (See Figure 6.1).

A number of previous studies have identified ohnologs in the human genome [Dehal and Boore, 2005; Putnam et al., 2008; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010]. However, only [Putnam et al., 2008] used genome-scale comparison with a reference genome (Amphioxus), diverged before 2R-WGD. Other studies only performed comparison of the human genome with itself. [Makino and McLysaght, 2010] and [Huminiecki and Heldin, 2010] adapted the approach described in [Dehal and Boore, 2005], using only a single window (100 genes). If there were at least two paralogs that were duplicated at the base of vertebrates in this window, the paralogs were taken as ohnolog candidates. More importantly, there has not been any quantitative measure to detect if the two paralogs can be identified on any window *by-chance*, or to give more weight to the windows with many paralogs. In this chapter, I will discuss the results of our approach, and how combining information from multiple outgroups can indeed improve our confidence in the identified candidates using the human genome as reference. The results for other vertebrates follow exactly the same trend and will be discussed in later sections.

9.1 Combining information from Multiple outgroups improves ohnolog detection

Table 9.1 lists the numbers of ohnologs identified by all six outgroups and Human-Human synteny comparison without any filtering on *P-value* or duplication time for outgroups. For self-comparison however, only paralogs duplicated at the base of vertebrate were considered. As we used the most relaxed criteria possible, this data is ridden with false positive candidates, either duplicated before or after WGD, or having very high *P-values*.

Table 9.1: Number of **Human Ohnologs** Identified by Outgroup and Self Comparison without any filter on *P-value* and duplication time

Comparison		Ohnolog Pairs for Window Size					Total Ohno Pairs
		100	200	300	400	500	
Human	*Amphioxus*	2012	4227	5930	7401	8368	8596
Human	*Ciona intestinalis*	4647	18318	27457	33994	39812	40589
Human	*Ciona savignyi*	7200	36130	47056	51144	52026	56521
Human	*Drosophila*	5843	32102	59613	62622	62624	63658
Human	*Sea Urchin*	1603	2106	2389	2685	3156	3337
Human	*Worm*	4367	30474	86869	104086	105079	106085
Human	*Human*	10200	12757	14041	14344	14204	15054
					Total Ohnologs		**165779**

Therefore, from the 165,779 combined ohnologs first we removed ohnologs not duplicated at the base of vertebrates for all outgroups leading to 15,107 ohnolog pairs (only 9.1 % "true" candidates) with *P-values* ranging from 3.1×10^{-20} to 0.9999. Of these 15,107 ohnolog pairs 11,428 were identified with at least one outgroup and 15,054 were identified by self comparison alone. We then constructed three set of ohnologs with increasing confidence by combining *P-values* from both outgroups and self-comparison (Section 6.8), leading to three sets of true positive ohnolog pairs:

- **Strict Criteria: 2,935 ohnolog pairs**
- **Intermediate Criteria: 5,119 ohnolog pairs**
- **Relaxed Criteria: 7,270 ohnolog pairs**

Otherwise noted, the ohnolog pairs characterized in this thesis refer to the 7,270 pairs from the Relaxed Criteria. To emphasize the importance of using multiple outgroups to detect ohnologs, a six way Edwards' Venn diagram (Figure 9.1) depicts the distribution of ohnolog pairs with respect to all outgroups. Ohnolog pairs range from 1,417 with sea urchin comparison to a maximum of 4,983 using the Worm, *C.elegans* as outgroup. There are only 290 ohnolog pairs identified by all outgroups, and a minimum of 27 (for Sea urchin) to 246 (for *C. intestinalis*) ohnolog pairs are identified by a single outgroup alone. To identify ohnologs in the vertebrate genomes, only the study by [Putnam et al., 2008] used comparison of human and a single reference genome of Amphioxus on a genome wide scale. Although a slightly different approach was utilized to perform macro-synteny comparison based upon construction of ancient chordate linkage group, where the window size may cover entire human chromosomes. However, as depicted in Figure 9.1, many ohnologs would not be identified using just a single outgroup genome owing to lineage specific rearrangements in the outgroup genomes, limitations of genome assembly/annotation or homology criteria.

9.1.1 Comparison with randomized human genome

Although it is clear that multiple outgroups lead to more ohnolog pairs. Yet, to make sure that our approach to combine *P-values* from the outgroups by multiplication does not include noise in the ohnologs in the form of false positives, we compared *P-values* from comparison by original and shuffled human genome. We randomized the human genome as described in

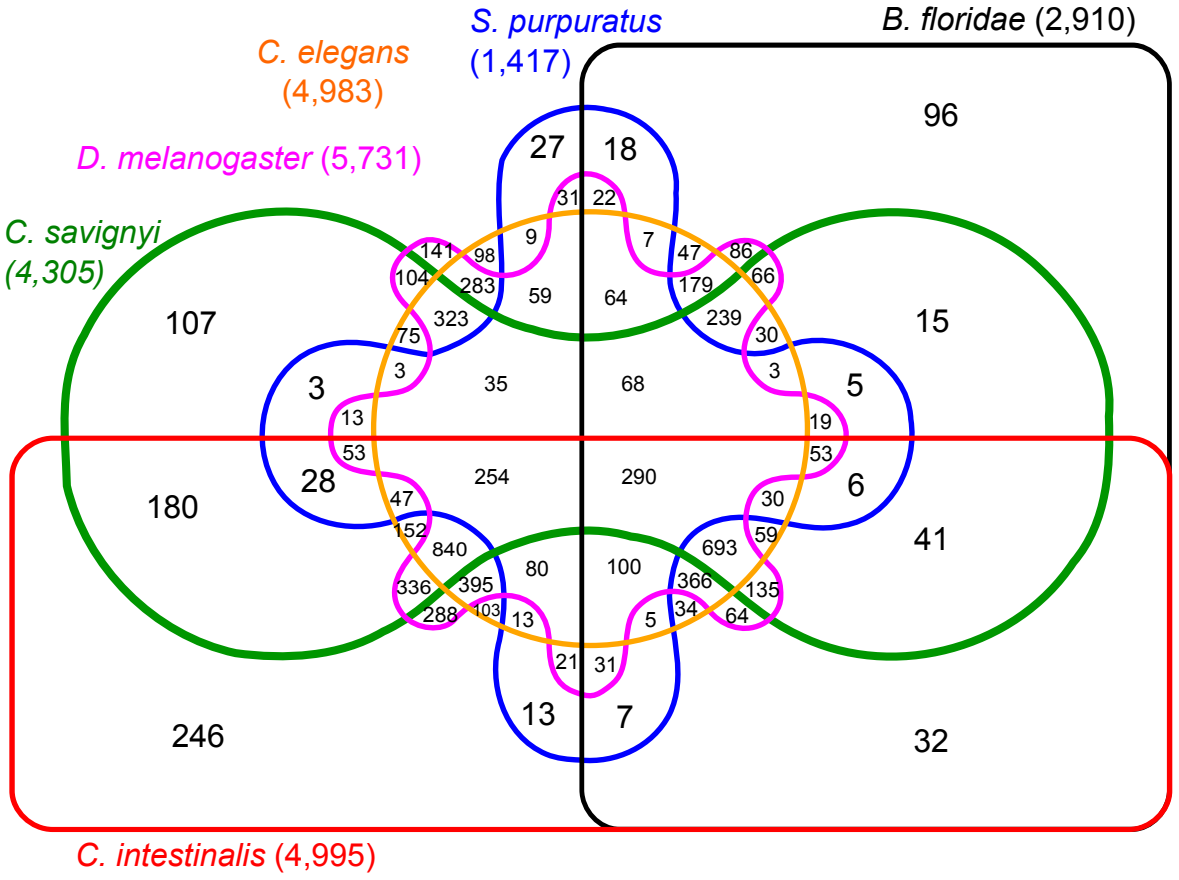

Figure 9.1: A six-way Venn diagram showing distribution of 7,270 human ohnolog pairs from relaxed criteria *wrt* all outgroups

Section 6.10. We then identified ohnologs using the window size of 300 using our approach by all the outgroups, in randomized and original genome. Finally we combined the *P-values* by multiplication for both the cases.

To avoid any bias due to number of identified ohnologs from comparison by original or random genomes, we assigned a probability value of 1 to any paralog pair which shares the same outgroup ortholog, is duplicated at the correct time, yet not identified in any of the two comparisons. It implies that the probability of these pairs not being ohnologs is one. This equalizes the total number of possible ohnologs (14,966).

The distribution of combined *P-values* (25 bins) for both of the scenarios is depicted in (Figure 9.2). While we can observed a sharp increase of the genes belonging to lowest probability bins for original human genome, for the randomized human genome considerably less genes belong to the low probability classes. If we remove the two basal invertebrates, which are expected to have least conserved synteny owing to the long time period of independent evolution since the divergence of vertebrate ancestors, we observed that there is hardly any enrichment of genes in the lower probability bins, as opposed to sharp increase observed for the actual genomes. These observations imply that our *P-value* based approach indeed does not incorporate much noise in the ohnolog dataset.

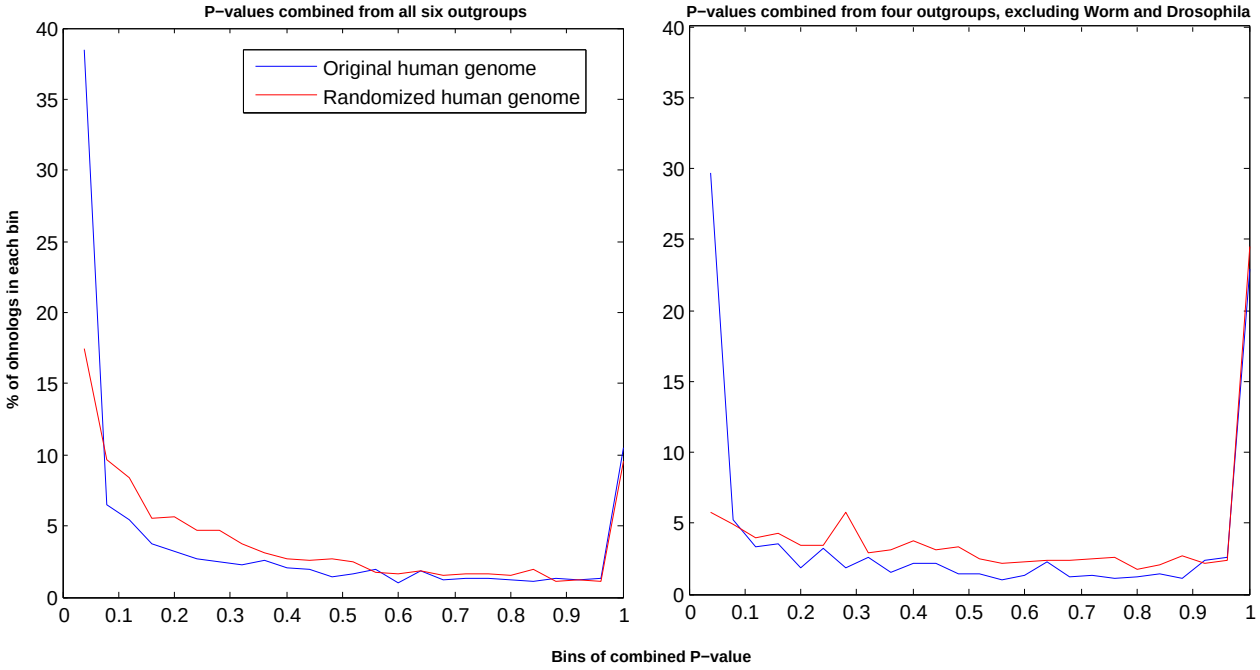

Figure 9.2: Comparisons of combined *P-value* distribution from original and randomized genomes

9.2 Comparison of ohnologs with published datasets

We then performed comparison of our ohnolog pairs with ohnolog pairs available from two published ohnolog datasets by [Makino and McLysaght, 2010] and [Huminiecki and Heldin, 2010]. [Makino and McLysaght, 2010] have identified 9,057 ohnolog pairs using comparison of human genome with itself with a window size of 100, having at least 2 ohnolog candidates in the window. Similarly, [Huminiecki and Heldin, 2010] also used human-human synteny comparison using i-ADHoRe 2.4 [Simillion et al., 2008].

9,057 ohnolog pairs from Makino and McLysaght were mapped on Ensembl version 70 leading to 8,378 pairs having both the ohnolog genes in version 70. Makino and McLysaght used sequence clustering to generate gene families using each of the six outgroup genome as reference one by one. Within these families, BLAST scores were used to identify candidates duplicated before fish tetrapod split. Finally ohnologs were combined from all the six outgroups. In our initial study [Singh et al., 2012], we have used ohnologs from Makino and Mclysaght and associated a *confidence index* with each ohnolog in terms of number of outgroups supporting the duplication time of a pair (1 to 6). Note however, that this outgroup support does not represent syntenic support unlike in the present method, rather it just represents the confidence that a paralog has been duplicated at the base of vertebrates.

Out of the 8,378 pairs from Makino and McLysaght, 5,448 also have correct duplication timing according to reconciled paralogs from Ensembl *v-70* (Section 6.1.2). Of these only 11 pairs could not be identified using our approach from self-comparison and 286 from outgroup comparison. In total, only 7 ohnolog pairs with correct duplication timing from Makino and McLysaght ohnologs dataset are not covered by both self and outgroup synteny comparison without any filter on *P-value*. This shows that our approach identifies most 'true' ohnologs. However, we suspect that many ohnolog candidates from Makino and Mclysaght are in fact likely false-positives.

For example, of the 2,930/8,378 ohnologs from Makino and McLysaght that are not dupli-

cated at the base of vertebrates in our dataset, 79% (2,310) pairs have a *confidence index* of only 1, 16% (475 pairs) of 2 or 3, and only 5% had the confidence index of more than half of [Makino and McLysaght, 2010] ohnologs [Singh et al., 2012]. These observations suggest that although they lie on a macro-synteny window, many of these pairs are not true positive candidates. In our outgroup comparison, since the orthologs are not filtered initially based on the correct duplication time, it was observed that 2,070/2,930 ohnolog pairs could be identified with at least one outgroup comparison. However, the average *P-value* of these 2,070 ohnolog pairs was 0.2, as opposed to the average *P-value* of 0.006 for ohnologs from the Relaxed criteria from our method. Yet, 223 of these 2,070 pairs have a combined *P-values* less than 0.001 from the outgroup comparisons. Moreover, 455 pairs have been identified by synteny comparison by more than half of the outgroups, suggesting that these pairs may likely be 'true' ohnologs. Therefore, the accurate identification of duplication time is also critical in the identification of ohnologs.

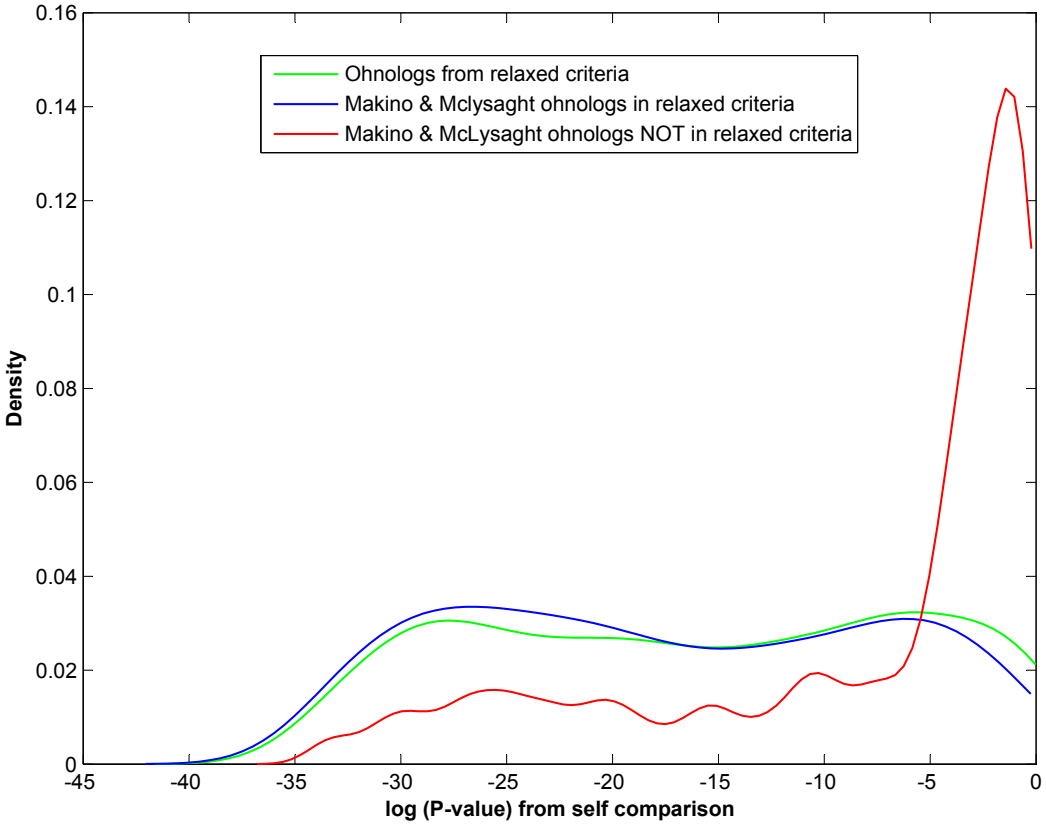

Figure 9.3: Density distribution of *P-values* from self comparison for Makino & McLysaght ohnolog pairs that are included or excluded as ohnologs with our Relaxed criteria

23%, 33% and 40% of the pairs respectively from strict, intermediate and relaxed criteria from our analysis are not part of ohnologs in Makino and McLysaght dataset. Figure 9.3 shows the density distribution of *P-values* from human-human comparison for Makino and McLysaght ohnologs that are identified by our relaxed criteria, and the ones that are excluded as ohnologs due to our *P-value* filters. Note that while *P-value* distribution of 4,395 Makino and McLysaght ohnologs is highly similar to overall distribution of 7,270 pairs from the Relaxed criteria, high peak at greater *P-value* for 1,042 ohnolog pairs that are not part of Relaxed criteria strongly suggests that most of these are false positive candidates.

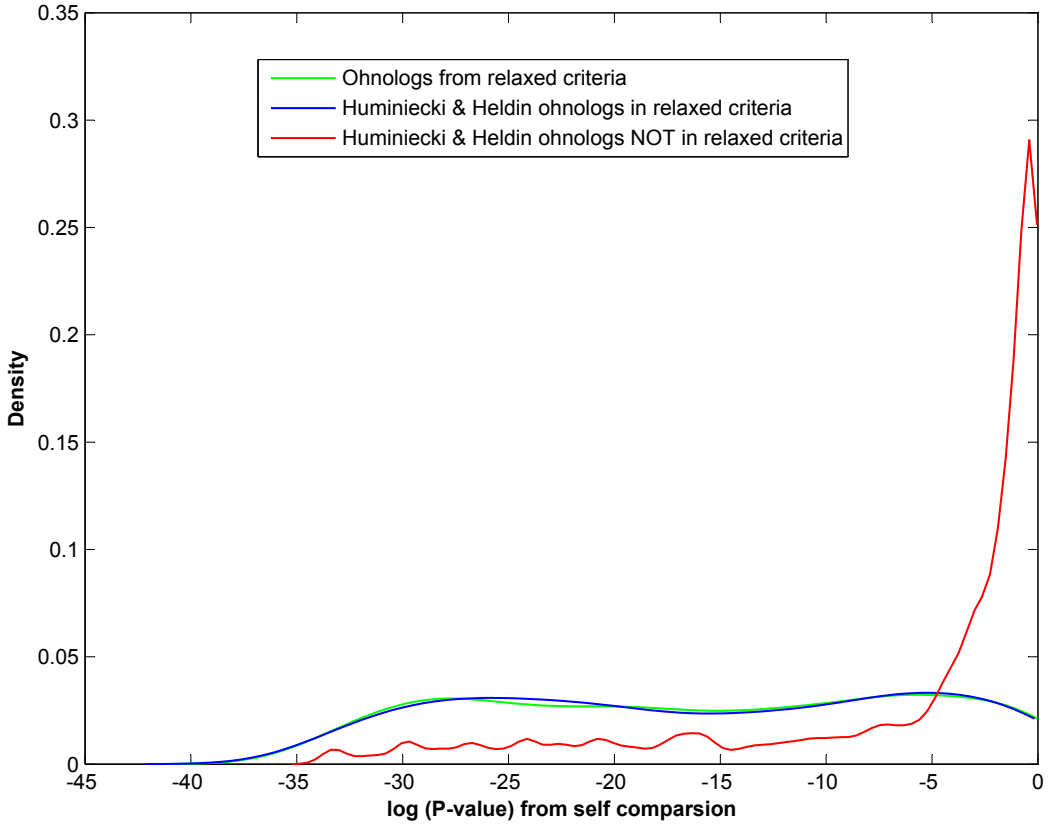

Figure 9.4: Density distribution of *P-values* from Huminiecki & Heldin ohnolog pairs that are included or excluded as ohnologs with our Relaxed criteria

We also obtained 56,053 ohnolog pairs from [Huminiecki and Heldin, 2010] who used duplication time from TreeFam database, based on reconciliation of gene trees in TreeFam with species tree [Li et al., 2006]. *Saccharomyces cerevisiae* and *Arabidopsis thaliana* have been used as the outgroups to root these trees. Out of these pairs 12,939 were redundant, where directionality of the homolog relation was not treated properly *e.g.* *RPS6KB1–RPS6KA3* and *RPS6KA3–RPS6KB1* were taken to be two different pairs. We could map 29,344 non-redundant pairs on Ensembl *v70*. Of these only 11,397 are duplicated at the base of vertebrates according to Ensembl gene trees. Without any filter on *P-value* we could identify 8,895 pairs using self-comparison and 7,578 using outgroup comparison. A total of 2,471 pairs could not be identified from both outgroup and self comparison. Most of these un-identified pairs belong to large paralogous families with high sequence similarity *e.g.* chemokines, interleukins and protocadherins. A distribution of *P-values* (self-comparison) from our Relaxed criteria and for the Huminiecki and Heldin ohnologs that are covered (5,295) or excluded (3,600) by our analysis is shown in Figure 9.4. Similar to Makino and McLysaght ohnologs we observe that for the ohnolog pairs excluded by our approach, the peak at higher *P-value* is even higher. These observations suggest that our approach to combine ohnolog pairs from multiple windows and more importantly filtering based on *P-value* can indeed improve the detection of true positive candidates and exclude ohnologs with low confidence.

In summary, while the approach of [Makino and McLysaght, 2010] works reasonably well, [Huminiecki and Heldin, 2010] dataset has much more false positives. In particular both the studies incorrectly includes genes duplicated much before the origin of vertebrates as ohnolog candidates. Owing to a very relaxed synteny criteria most of these are identified as ohnologs.

While a possibility of whole genome duplication before 2R-WGD can not be ruled out, these pairs are certainly not ohnologs from the 2R-WGD.

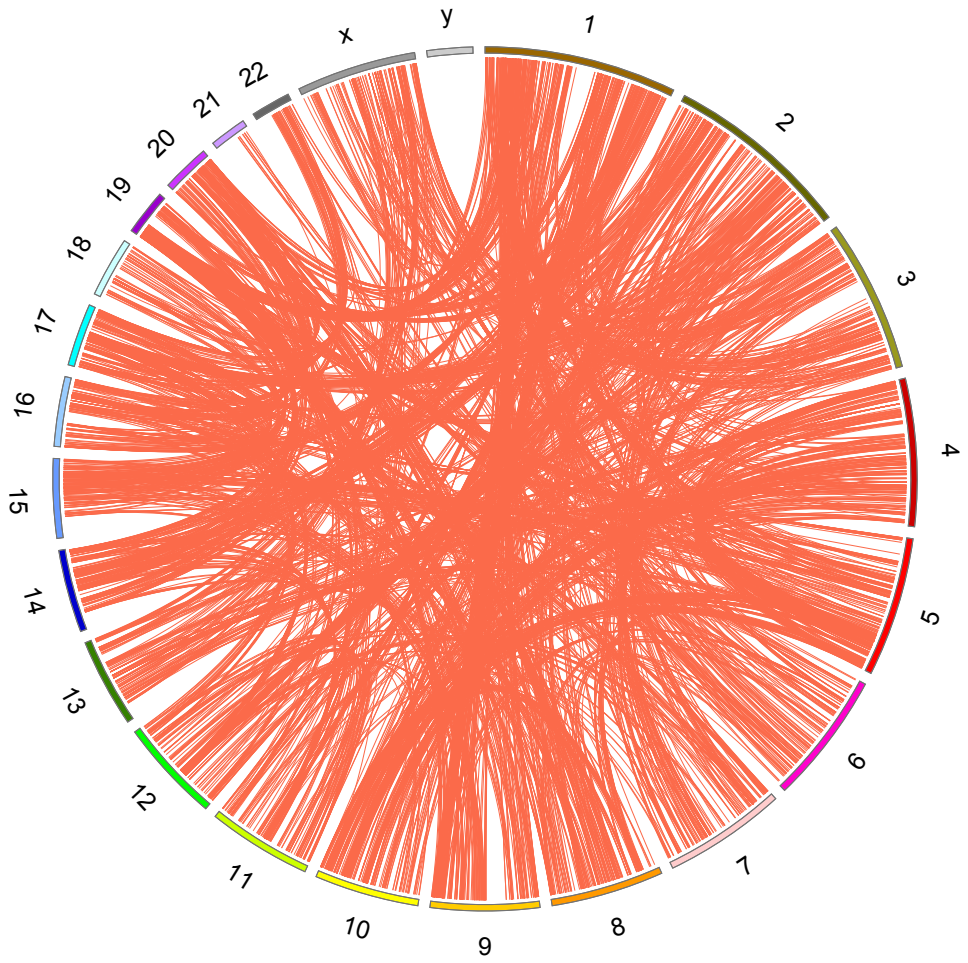

Figure 9.5: Genomic distribution of human ohnologs identified by all the three criteria

9.3 Ohnolog family size distribution

Using the ohnolog pairs we next set to construct paralogous families retained from 2R-WGD. If a gene retains all ohnolog partners, the maximum family size is expected to be four. However, as most genes loose their partners after duplication, most ohnolog families should belong to size two or three.

We could generate 1,473, 2,102 and 2,551 ohnolog families from strict, intermediate and relaxed filters respectively for the human genome. Most remarkably, for almost all of these families, the size never exceeds mores than four, as expected for the two rounds of genome duplications. As depicted in Table 9.2, all but 7 ohnolog families belong to a size of at least four with strict filters on *P-value*. Even with the most relaxed *P-values* filters, 96.6% ohnolog families have only 2, 3 or 4 ohnolog partners. Furthermore, a sharp decline in the number of families could be observed beyond size four, another strong sign of two rounds of genome duplications.

We also used our approach to generate the ohnolog families from the pairs provided by [Makino and McLysaght, 2010] and [Huminiecki and Heldin, 2010]. The number of families with size upto four all pairs from Makino & McLysaght was found to be 93.8%, lower than from

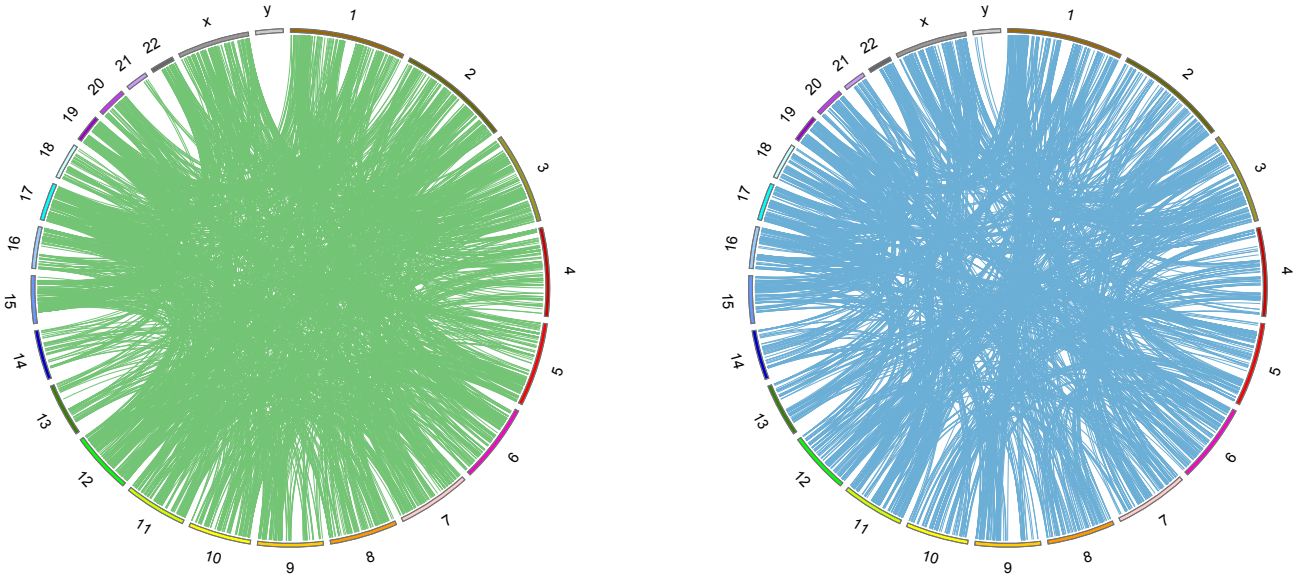

Figure 9.6: Genomic distribution of human ohnologs identified by intermediate (left) and only relaxed criteria (right)

Table 9.2: Individual ohnologs, pairs and families for different criteria in the human genome

Criteria	Ohno Pairs	Individual Ohnologs	Ohnolog Families	Family Sizes 2	3	4	> 4	% of families with size ≤ 4
Ohnologs from our study								
Strict	2935	3783	1473	1011	365	89	7	99.5%
Intermediate	5733	5719	2102	1393	500	162	46	97.7%
Relaxed	7270	7351	2551	1644	601	219	82	96.6%
Ohnologs from [Makino and McLysaght, 2010]								
All pairs	8378	6989	2352	1466	543	199	97	93.8%
Conf. Ind. ≤ 3	5347	4116	1267	792	260	92	79	90.2%
Conf. Ind. ≥ 4	3031	3923	1582	1099	364	102	11	98.7%
Ohnologs from [Huminiecki and Heldin, 2010]								
All pairs	29344	9557	2600	1222	618	332	348	86.6%

the relaxed criteria in this study. This fraction becomes even lower for the ohnolog pairs with *confidence index* [Singh et al., 2012] less than four (90.2%). Only 'high quality' ohnolog pairs with *confidence index* greater than three have 98% gene families within size four, comparable to our approach. Ohnolog pairs from Huminiechi & Heldin performed even worse with only 86.6% ohnolog families belonging to a size of 2, 3 or 4.

9.4 Ohnolog pairs for other vertebrates

In addition to the human genome, we also identified ohnologs in five other vertebrate genomes: chicken, dog, mouse, pig and rat. The same approach and the six outgroups were used to compute ohnolog pairs and generated ohnolog families. Table 9.3 shows a summary of individual ohnologs, pairs and ohnolog families for these genomes using strict, intermediate and relaxed *P-value* criteria.

Table 9.3: Individual ohnologs, pairs and families for all the three criteria in five vertebrate genomes

Organism	Criteria	Ohno Pairs	Individual Ohnologs	Ohnolog Families	Family Sizes 2	3	4	> 4	% of families with size ≤ 4
	Strict	1286	2010	890	752	122	15	1	99.9%
	Intermediate	1895	2809	1210	974	204	28	4	99.7%
	Relaxed	3030	4214	1753	1356	330	57	10	99.4%
	Strict	924	1432	610	474	117	18	1	99.8%
	Intermediate	2257	3098	1266	926	271	61	8	99.4%
	Relaxed	3940	4893	1878	1285	448	124	21	98.9%
	Strict	2495	3225	1262	917	260	73	12	99.0%
	Intermediate	5701	5809	2085	1309	534	178	63	96.9%
	Relaxed	8315	7007	2413	1487	600	226	95	95.7%
	Strict	1911	2597	1005	768	191	42	4	99.6%
	Intermediate	3801	4470	1609	1114	350	108	37	97.7%
	Relaxed	5496	5886	1999	1342	451	140	66	96.7%
	Strict	2698	3472	1320	934	280	87	19	98.6%
	Intermediate	6623	6482	2164	1286	552	221	102	95.2%
	Relaxed	8811	7693	2451	1411	645	245	141	93.9%

The level of annotation of these genomes is variable and annotated protein coding genes also range from 15,310 for chicken to 22,865 for the rate genome. Using relaxed *P-value* filters a minimum of 3,030 to a maximum of 8,811 ohnolog pairs could be identified for chicken and rat respectively. Ohnolog family size consistently follows the same distribution where there is a sharp decline in the number of families after size four and, almost all families belong to size 2, 3 or 4. Remarkable, for the chicken, only 1, 4 and 10 ohnolog families belongs to a size more than 4 with strict, intermediate and relaxed criteria respectively. This might, in part reflect the slow rate of rearrangements in the chicken lineage [Hillier et al., 2004]. However, the total annotated protein coding genes in chicken are highly underestimated (only 15,310) by current annotations in Ensembl. Therefore, the total number ohnologs are also less as compared to other analysed vertebrates. In fact, we suspect that the differences in the number of ohnologs and families in different vertebrates is primarily due to limited genome annotation or discrepancies in the ortholog or paralog dataset, rather than lineage specific gene loss.

9.5 The OHNOLOGS server

The data of all the ohnolog pairs and families for the six tetrapod genomes has been made accessible on the web sever: `http://ohnologs.curie.fr/`. Using *OHNOLOGS*, users can search for a particular gene browse pre-compiled ohnolog families and pairs or generate ohnolog families based on P-value filters.

9.5.1 Search

On the *Search* page (Figure 9.7), the user can search for the gene of their interest in any of the six vertebrates using either Ensembl Id, gene symbol or any desired keywords. A keyword search, if it does not match any gene symbol directly, displays all the genes matching that keyword in gene symbol or description. A hyperlink from this page directs to the page having

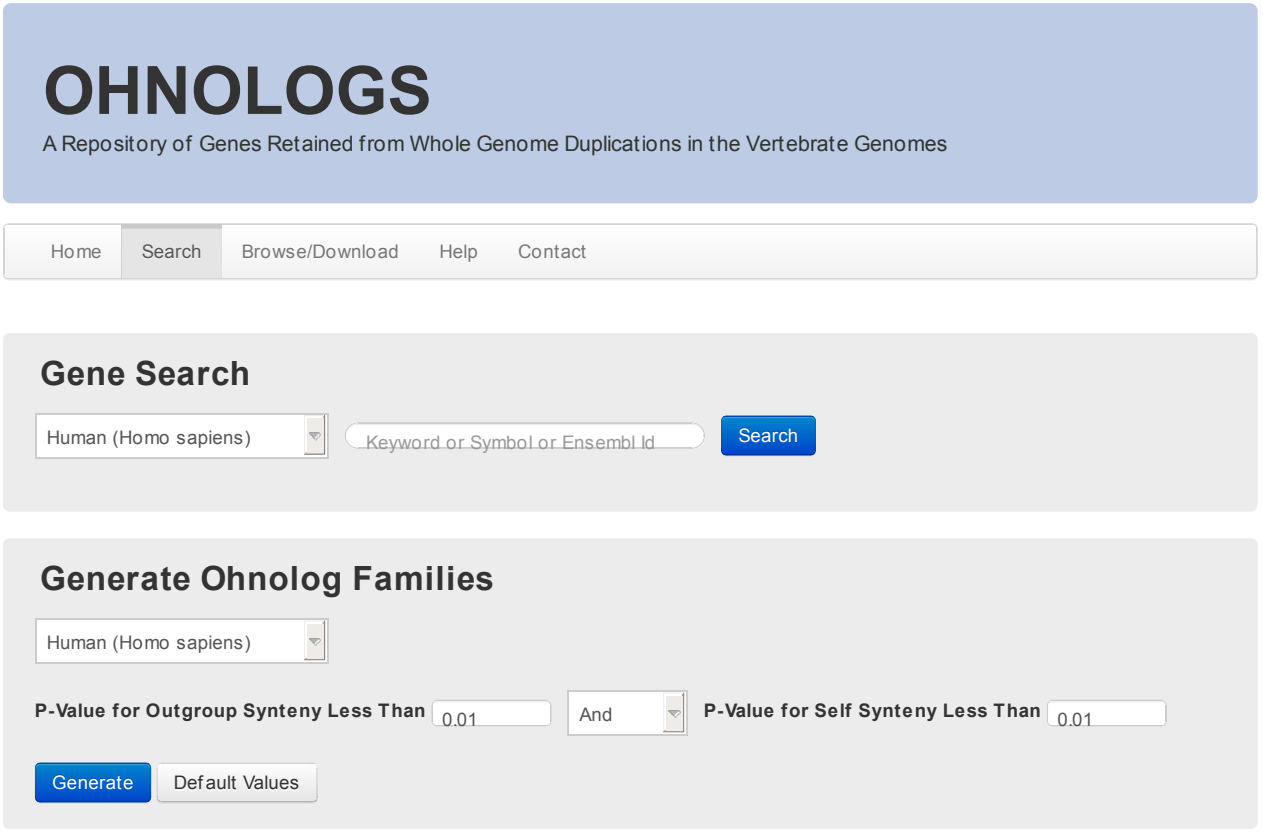

Figure 9.7: Search Page on the *OHNOLOGS* server

details of ohnolog families. If the gene exists in our analysis, and is an ohnolog, user will directly be directed to the details about ohnolog families.

On the search page user can also generate ohnolog families using our approach, for any of the six vertebrate genome using a desired *P-value* criteria for outgroup or self comparison. The default values are from the Strict criteria. The results page displayed all the generated families, which can also be downloaded easily. To implement this feature, all the 165,779 ohnolog pairs are taken as input and pairs are filtered based on the criteria combination provided by the user. Followed by a depth first search run, locally on the server to generate desired families.

9.5.2 Interpretation of an ohnolog family

The result page for ohnolog family search for human Ubiquilin genes is depicted in Figure 9.8. Families from all the three *P-value* criteria are displayed. using the strict criteria, only the pair *UBQLN1 – UBQLN4* can be identified as ohnologs. Relaxing the *P-value* results in a family of size three. Ohnolog partners for the families are displayed in different columns. Genes within the same cell are small scale duplicates e.g. *UBQLN1 – UBQLN2* and *UBQLN3 – UBQLBNL*. We use two different separators for SSDs: a comma (,) to distinguish if it is a recent SSD (after 2R-WGD), and a pipe (|) for an ancient SSD (before 2R-WGD). Hence, *UBQLN3 – UBQLBNL* have been duplicated by an SSD before the 2R-WGD, while *UBQLN1 – UBQLN2* have been duplicated by a recent SSD. It implies that the entire region having *UBQLN3 – UBQLBNL* genes was duplicated by the genome duplications. Duplication time are taken from Ensembl Compara.

As described in Section 6.9, if the SSDs do not lie within the same window, we also attempt

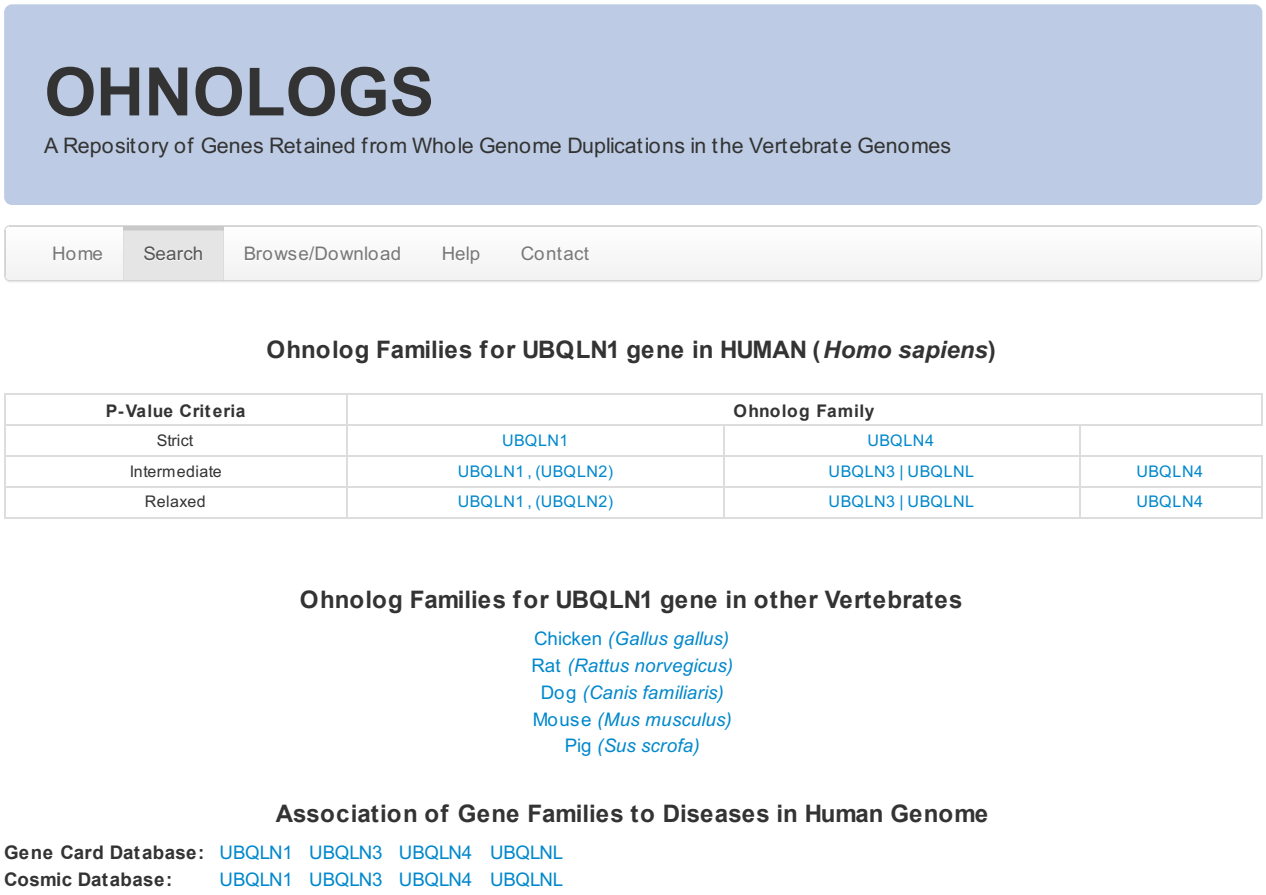

Figure 9.8: Ohnolog Family Page on the *OHNOLOGS* server, for Ubiquilin family for human.

to identify the region which might actually have been duplicated by WGD. In such instances, the ohnolog(s) on the region having strong synteny are kept outside the brackets e.g. *UBQLN1*, and the ones with weaker synteny are in the brackets, e.g. *UBQLN2*.

Clicking on the gene symbols directs to Ohnog Pairs page, where the pairs used to construct the family and the *P-values* for all the comparisons can be viewed.

A link to ohnolog family for the searched gene for other vertebrates has also been provided, along with the disease association (from COSMIC) and the details of the genes (from GeneCards database).

9.5.3 Browse & Download

We have also kept pre-calculated ohnolog pairs and families for all the vertebrates on the server. Using the Browse facility, users can view the ohnologs with the details of outgroups identifying them and the *P-values* for all comparisons. These pairs can also be downloaded for custom analysis. Similarly, ohnolog families from all the three *P-value* criteria can also be explored or downloaded from the *OHNOLOGS* server.

9.6 Ohnologs in the Teleost fish genomes

Teleost fishes are arguably the most species rich vertebrate group, having almost half of all vertebrate species [Taylor et al., 2003; Santini et al., 2009]. More importantly, they hold great potential to understand vertebrate genome evolution, due to an additional WGD (the 3R-WGD) that occurred after the divergence of teleost lineage from tetrapods [Amores et al., 1998; Jaillon et al., 2004]. Many of the teleost fish genomes have been sequenced and are phylogenetically uniquely poised to understand the impact of genome duplications on the evolution of vertebrates. We therefore selected four teleost fish genomes: Medaka (*Oryzias latipes*), Stickleback (*Gasterosteus aculeatus*), Tetraodon (*Tetraodon nigroviridis*) and Zebrafish (*Danio rerio*), where the chromosome level assemblies were available. We used our approach to calculate ohnologs retained from both the 2R and the 3R-WGD in all of the four fish genomes.

9.6.1 Ohnologs from the 2R-WGD

To identify the ohnologs retained from the vertebrate ancestral WGDs, we used the same six invertebrate genomes as outgroups. Table 9.4 lists the numbers of identified ohnolog pairs and families. Similar to the tetrapod genomes, we observed that most of the families belong to a size two or three, with very few above size four. The numbers of total ohnolog pairs identified are also comparable to the tetrapod genomes. The large number of pairs observed in Zebrafish is primarily due to the large number of annotated protein coding genes and total paralog pairs. Yet, with the relaxed *P-value* criteria, 95% of all 2R-ohnolog families are below size five. The results from the strict *P-value* filters are highly consistent for all the four genomes with close to 99% families below size five for Medaka and Stickleback genomes. Best results were obtained from the Tetraodon genome where even with the relaxed criteria almost all families belong to size of 2, 3 or 4. These results further validate that our approach works very well, even for a relatively recent WGD.

Table 9.4: Individual ohnologs, pairs and families **from the 2R-WGD** for all the three criteria in Medaka, Stickleback, Tetraodon and Zebrafish genomes

Organism	Criteria	Ohno Pairs	Individual Ohnologs	Ohnolog Families	Family Sizes				% of families with size ≤ 4
					2	3	4	> 4	
	Strict	2377	2997	1095	819	209	54	13	98.8%
	Intermediate	6489	5865	1886	1269	432	137	48	97.5%
	Relaxed	8562	6974	2132	1407	492	159	74	96.5%
	Strict	3373	3205	1109	842	191	59	17	98.5%
	Intermediate	8787	6693	2064	1359	495	144	66	96.8%
	Relaxed	11238	7788	2271	1454	555	162	100	95.6%
	Strict	1482	1929	733	590	119	20	4	99.5%
	Intermediate	3969	4172	1427	1050	281	79	17	98.8%
	Relaxed	5516	5266	1691	1194	351	107	39	98.4%
	Strict	4686	4378	1422	1042	268	87	25	98.2%
	Intermediate	14460	9006	2419	1567	545	200	107	95.6%
	Relaxed	19097	10455	2616	1643	604	234	135	94.8%

Table 9.5: Individual ohnologs, pairs and families **from the 3R-WGD** for all the three criteria in Medaka, Stickleback, Tetraodon and Zebrafish genomes

Organism	Criteria	Ohno Pairs	Individual Ohnologs	Ohnolog Families	Family Sizes 2	Family Sizes > 2	% of families with size ≤ 2
	Strict	1897	3621	1756	1748	8	99.5%
	Intermediate	2095	3903	1878	1864	14	99.3%
	Relaxed	2209	4019	1918	1903	15	99.2%
	Strict	2229	3934	1883	1874	9	99.5%
	Intermediate	2463	4234	2009	1993	16	99.2%
	Relaxed	2706	4350	2049	2032	17	99.2%
	Strict	1418	2644	1263	1260	3	99.8%
	Intermediate	1579	2936	1396	1389	7	99.5%
	Relaxed	1602	2979	1416	1409	7	99.5%
	[Jaillon et al., 2004]	1973	3647	1713	1687	26	98.4%
	Strict	4916	4725	2000	1989	11	99.5%
	Intermediate	4773	5516	2371	2345	26	98.9%
	Relaxed	5903	5855	2430	2399	31	98.7%

9.6.2 Ohnologs from the 3R-WGD

To identify ohnologs retained from the third fish specific WGD, we used Coelacanth along with the four tetrapods genomes as outgroups. The results are shown in Table 9.5, show remarkable consistency. Due to a single round of WGD, as expected, almost all of the ohnologs have just one more ohnolog partner. With the strict *P-value* filters, close to 100% ohnologs make a family of two ohnologs for all the four genomes. Even using the intermediate and relaxed filters, more than 99% ohnologs have just one more paralog.

We also compared the 3R-ohnolog pairs for *Tetraodon* identified from our approach to [Jaillon et al., 2004; Howe et al., 2013] [1]. We noticed that our approach cannot identify only 48/1,973 ohnologs by [Jaillon et al., 2004]. None of these 48 has been duplicated at the three nodes we took as the candidate ohnolog duplication time. However, after applying our *P-value* filters: 1,299, 1,413 and 1,429 ohnologs out of 1,973 were identified to be the true ohnologs. More interestingly, we noticed that 414/1,973 pairs in the [Jaillon et al., 2004] dataset were duplicated at the base of vertebrates, and were in fact identified as ohnologs from 2R-WGD in fish genome rather than 3R. All in all, our approach works very well for fish specific WGD and the ohnolog pairs have a very good correspondence with [Jaillon et al., 2004] dataset.

[1] Courtesy: Camille Berthelot and Hugues Roest-Crollius

10

Enhanced Retention of "Dangerous" Genes by WGD

JUST as some genes happen to be more "essential," owing to their deleterious loss-of-function or null mutations, some genes can be classified as more "dangerous," due to their propensity to acquire deleterious gain-of-function mutations. This is, in particular, the case for oncogenes, dominant disease genes and genes with autoinhibitory protein folds. Mutation in these genes typically lead to constitutively active mutants with dominant deleterious phenotypes, often detrimental for the life and fitness of organisms.

Disease genes in humans have been shown to be under strong purifying selection [Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009]. Yet, "dangerous" gene families implicated in cancer and severe genetic diseases have also been greatly expanded by duplication in the course of vertebrate evolution. The emergence and evolutionary expansion of these gene families is an evolutionary oddity from a natural selection perspective. In this chapter, I will characterize the mechanism of duplication of these genes, and shed some light on the evolutionary constraints that gave rise to their biased retention in the human genome.

To analyze a possible association between the susceptibility of human genes to deleterious mutations and their duplication mechanism, we considered multiple classes of genes susceptible to deleterious mutations from experimentally verified databases and literature. These classes include cancer genes, Mendelian disease genes in human, dominant negative genes and genes with autoinhibitory protein folds (Chapter 7). Our aim was to look at the relative contributions of WGD and SSD in the expansion of these "dangerous" gene classes.

10.1 The Majority of "dangerous" genes retain more ohnologs

We first investigated the retention of ohnologs among these "dangerous" genes classes. We overlaid the disease genes classes with our ohnologs from the Relaxed criteria. As depicted in Figure 10.1, there is indeed a strong association between the retention of human ohnologs from vertebrate WGD and their reported susceptibility to deleterious mutations.

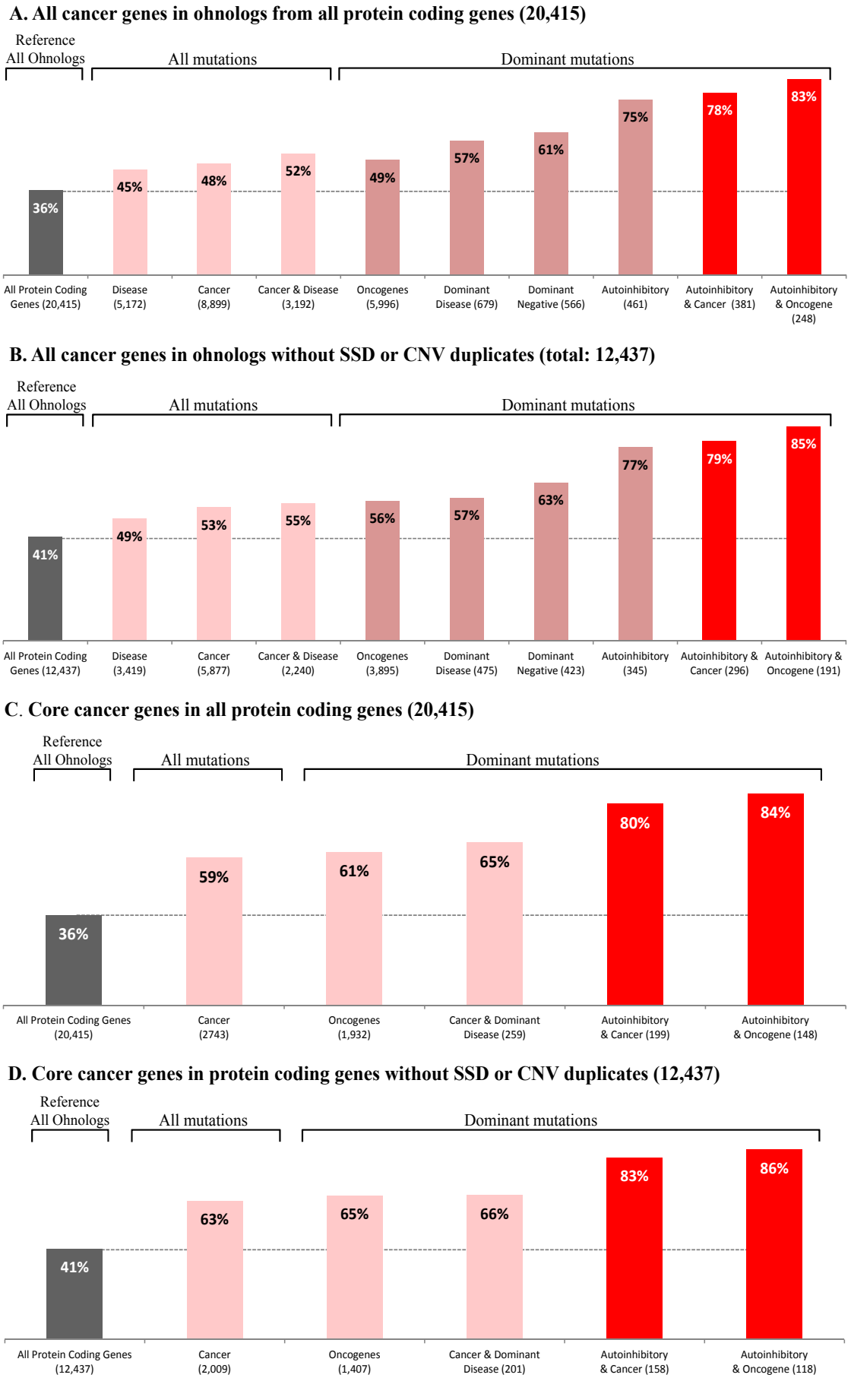

Figure 10.1: Prevalence of Human Ohnologs in Gene Classes Prone to Deleterious Mutations

All gene classes susceptible to mutations have retained significantly more ohnologs than expected by chance. Overall, 45% of Mendelian disease genes in the human genome belong to ohnologs as opposed to the expected genome distribution of ohnologs *i.e.* 36% (45%; 2,322/5,172; p-value = $1.879 \times 10^{-15}, \chi^2$ test) (Figure 10.1 A). Similarly, all cancer associated genes in the human genome have also retained significantly higher ohnologs (48%; 4,236/8,899; p-value = $6.806 \times 10^{-115}, \chi^2$ test). Furthermore, these associations, which do not take into account the actual severity of the gene mutations are clearly enhanced when the analysis is restricted to genes with dominant deleterious mutations, such as oncogenes (49%; 2,941/5,996; p-value = $2.966 \times 10^{-98}, \chi^2$ test), dominant disease genes (57%; 389/679; p-value = $7.136 \times 10^{-31}, \chi^2$ test), dominant negative genes (61%; 345/566; p-value = $4.104 \times 10^{-35}, \chi^2$ test) and autoinhibitory genes (75%; 345/461; p-value = $1.441 \times 10^{-67}, \chi^2$ test). The biased retention of ohnologs is even stronger for genes combining several factors associated with an enhanced susceptibility to deleterious mutations, such as Mendelian disease genes mutated in cancer (52%; 1,654/3,192; p-value = $2.812 \times 10^{-77}, \chi^2$ test), cancer genes with autoinhibitory folds (78%; 296/381; p-value = $1.944 \times 10^{-64}, \chi^2$ test), or oncogenes with autoinhibitory folds (83%; 207/248; p-value = $1.162 \times 10^{-54}, \chi^2$ test).

Ohnologs and small scale duplications (SSD) are known to have antagonistic retention patterns [Davis and Petrov, 2005; Hakes et al., 2007; Fares et al., 2013]. It has been shown that Ohnologs in the human genome are also refectory to copy number variations (CNV) [Makino and McLysaght, 2010]. Therefore, to remove any bias from SSD or CNV, we next restricted ourself to ohnologs that have not experienced any recent SSD after WGD, and that do not have variable copy numbers in the human genome (12,437 genes after removing SSD & CNV genes). We were again able to observe an even more pronounced retention of ohnologs without SSD nor CNV in dangerous genes (Figure 10.1 B). Both Mendelian disease genes (49%; 1,661/3,419; p-value = $5.519 \times 10^{-18}, \chi^2$ test) and cancer genes (53%; 3,123/5,877; p-value = $7.993 \times 10^{-76}, \chi^2$ test) have retained even greater number of ohnologs without SSD/CNV. Similar to our observations on all protein coding genes, oncogenes (56%; 2,169/3,895; p-value = $2.973 \times 10^{-74}, \chi^2$ test), dominant disease genes (57%; 270/475; p-value = $6.085 \times 10^{-12}, \chi^2$ test), dominant negative genes (63%; 266/423; p-value = $1.985 \times 10^{-19}, \chi^2$ test), and autoinhibitory genes (77%; 266/345; p-value = $1.483 \times 10^{-41}, \chi^2$ test), have also retained significantly greater ohnologs. Cancer & disease genes (55%; 1,227/2,240; p-value = $2.345 \times 10^{-38}, \chi^2$ test), cancer and autoinhibitory genes and (79%; 235/296; p-value = $2.064 \times 10^{-40}, \chi^2$ test), autoinhibitory oncogenes (85%; 163/191; p-value = $4.291 \times 10^{-35}, \chi^2$ test), also show greater retention in ohnologs.

If we restrict ourself to cancer genes from core dataset, we expect that the disease association of ohnologs would be even stronger. Consistent with our hypothesis, (Figure 10.1 C & D) 2,743 genes from our core cancer dataset are highly enriched in ohnologs, with all genes, (59%; 1,612/2,43; p-value = $3.987 \times 10^{-136}, \chi^2$ test), and without SSD/CNV duplicates (63%; 1,269/2,009; p-value = $2.495 \times 10^{-88}, \chi^2$ test). Genes susceptible to dominant deleterious mutations, such as all oncogenes (61%; 1,180/1,932; p-value = $1.317 \times 10^{-116}, \chi^2$ test), and oncogenes without SSD/CNV (65%; 921/1,407; p-value = $8.928 \times 10^{-76}, \chi^2$ test), from core cancer genes also display stronger ohnolog associations. Similarly, focusing on genes belonging to multiple disease categories from core cancer genes, e.g. cancer and dominant disease (65%; 168/259; p-value = $3.860 \times 10^{-22}, \chi^2$ test), cancer genes with autoinhibitory folds (80%; 160/199; p-value = $6.662 \times 10^{-39}, \chi^2$ test), and oncogenes with autoinhibitory folds (84%; 125/148; p-value = $1.168 \times 10^{-34}, \chi^2$ test), stronger enrichment in ohnologs can be observed. In protein coding genes without recent SSD and CNV, this prevalence reaches up to 86% *vs* global ohnolog retention of 41% for 118 autoinhibitory-oncogenes from core dataset (86%; 102/118; p-value = $2.188 \times 10^{-23}, \chi^2$ test), with consistent pattern from, cancer and dominant disease genes (66%;

132/201; p-value = $2.158 \times 10^{-12}, \chi^2$ test), and cancer genes with autoinhibitory folds (83%; 131/158; p-value = $2.188 \times 10^{-26}, \chi^2$ test).

10.1.1 Ohnolog–disease association is consistent for high confidence ohnolog datasets

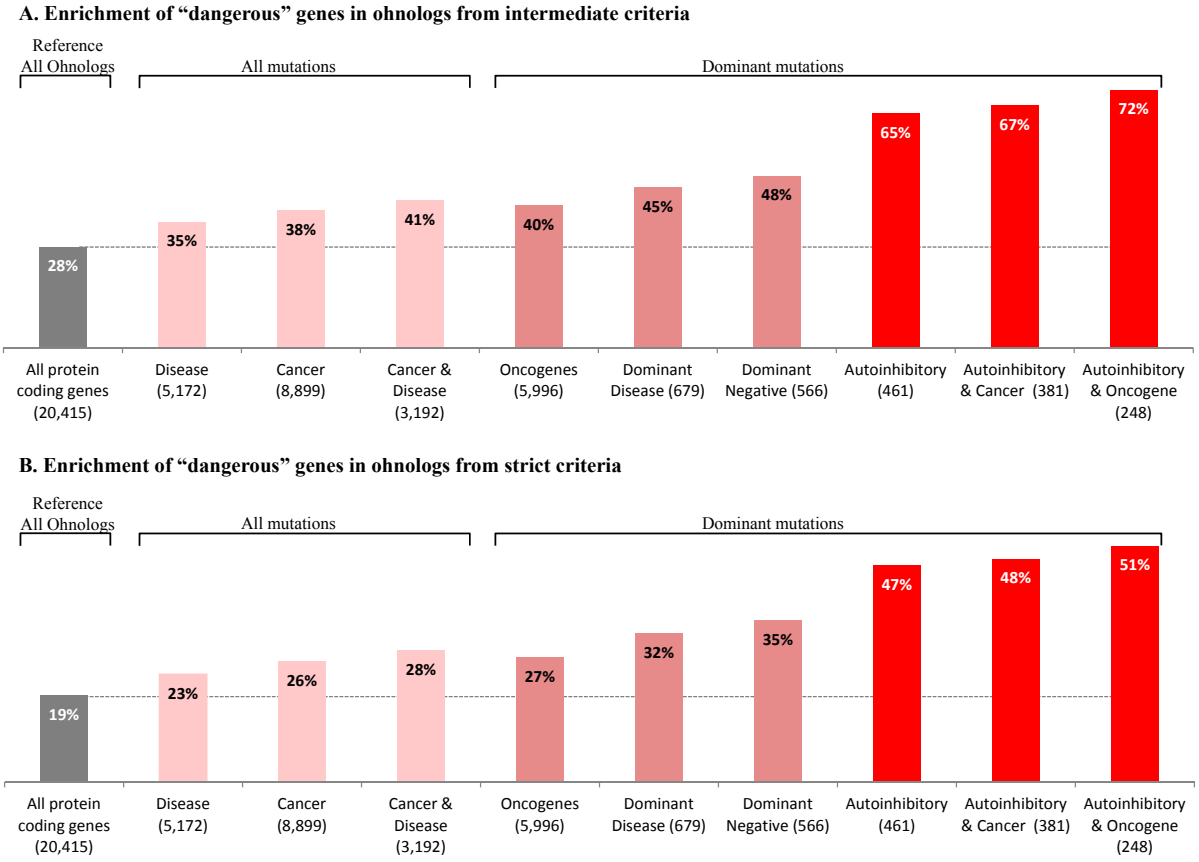

Figure 10.2: Prevalence of high confidence ohnologs in disease genes in the human genome

To test the robustness of our observations for disease genes' association with ohnologs, we restricted ourselves to ohnologs from intermediate and strict *P-value* criteria. We assumed that all the paralog pairs, not identified as ohnologs with the stricter criteria are in fact not ohnologs. Yet, we were able to confirm that indeed disease genes in the human genome have retained significantly more ohnologs than expected by chance (Figure 10.2).

Against the genome average of 28% (5,733/20,514) ohnologs from intermediate criteria, all the classes susceptible to mutations, e.g. disease genes (35%; 1,811/5,172; p-value = $1.326 \times 10^{-28}, \chi^2$ test), cancer genes (38%; 3,398/8,899; p-value = $8.625 \times 10^{-100}, \chi^2$ test), and both cancer & disease genes (41%; 1,310/3,192; p-value = $1.156 \times 10^{-59}, \chi^2$ test), have consistently retained more ohnologs with extremely lower χ^2 p-values. The retention bias is stronger for the genes susceptible to dominant mutations, with oncogenes (40%; 2,374/5,996; p-value = $1.527 \times 10^{-87}, \chi^2$ test), dominant disease genes (45%; 303/679; p-value = $8.669 \times 10^{-22}, \chi^2$ test), dominant negative genes (48%; 271/566; p-value = $1.060 \times 10^{-25}, \chi^2$ test), autoinhibitory genes (65%; 301/461; p-value = $1.046 \times 10^{-70}, \chi^2$ test), cancer genes with autoinhibitory folds (67%; 257/381; p-value = $1.466 \times 10^{-65}, \chi^2$ test), and oncogenes with autoinhibitory folds (72%; 178/248; p-value = $6.494 \times 10^{-53}, \chi^2$ test) have retaining significantly more ohnologs (Figure 10.2 A).

Even if we assume that all but 3,783 ohnologs from strict criteria are non-ohnologs, the stronger retention bias for "dangerous" genes persists. Against 19% genome average of ohnologs, mendelian disease genes (23%; 1,200/5,172; p-value = 5.318×10^{-18}, χ^2 test), cancer associated genes (26%; 2,307/8,899; p-value = 4.689×10^{-72}, χ^2 test), genes implicated in both cancer & other diseases (28%; 902/3,192; p-value = 2.009×10^{-45}, χ^2 test), oncogenes (27%; 1,603/5,996; p-value = 4.356×10^{-60}, χ^2 test), dominant disease genes (32%; 217/279; p-value = 2.144×10^{-19}, χ^2 test), dominant negative genes (35%; 197/566; p-value = 2.161×10^{-23}, χ^2 test), genes with autoinhibition (47%; 215/461; p-value = 2.106×10^{-54}, χ^2 test), and autoinhibitory genes implicated in cancer with all types of mutations (48%; 183/381; p-value = 1.082×10^{-49}, χ^2 test), or with dominant cancer mutations (51%; 126/248; p-value = 4.189×10^{-39}, χ^2 test), show strong bias towards ohnologs.

10.1.2 Enhanced retention of "dangerous" ohnologs in Mouse & Rat genomes

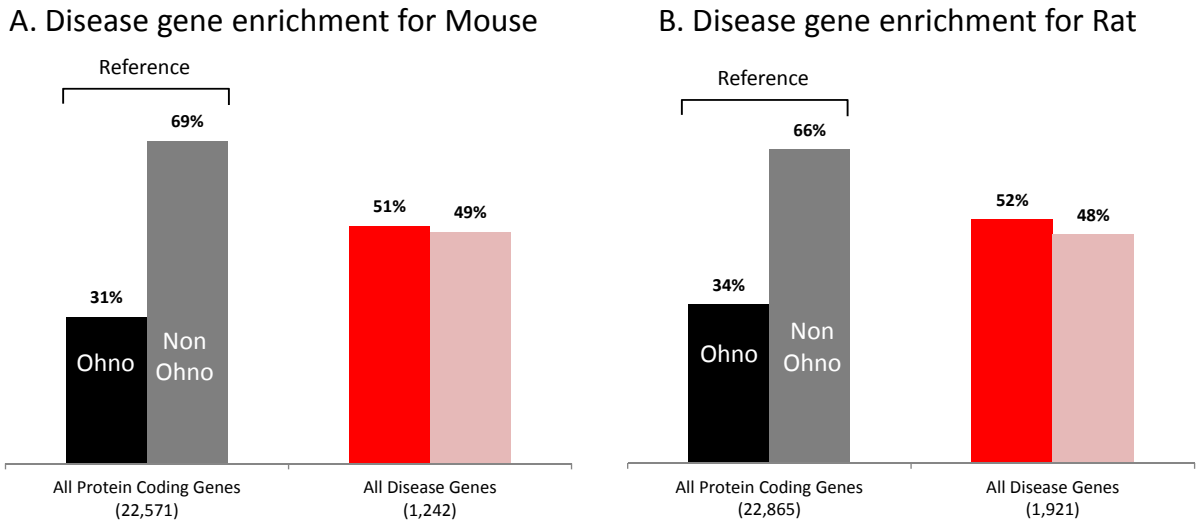

Figure 10.3: Prevalence of A. Mouse and B. Rat disease genes in ohnologs

We also investigated the prevalence of ohnologs in disease genes in the mouse and rat genomes. Since, mouse and rat are model organisms, we could obtain the information on 1,242 disease associated genes for mouse and 1,921 disease and cancer genes for rat, from Mouse Genome Informatics (MGI) [Eppig et al., 2012] and Rat Genome database (RGD) [Laulederkind et al., 2013] respectively.

We investigated the disease association of ohnologs for mouse and rat from relaxed *P-value* criteria. Consistent with our observations for the human ohnologs, we observe that disease associated genes in mouse have retained significantly more ohnologs 51% *vs* 31% (51%; 629/1,242; p-value = 4.525×10^{-49}, χ^2 test). Similarly cancer and disease genes in the rat genome have also retained more ohnologs than expected by chance, 52% *vs* 34% (52%; 992/1,921; p-value = 1.502×10^{-62}, χ^2 test).

All in all, these observations clearly suggest that human genes that are susceptible to deleterious mutations have retained a significant proportion of their ohnolog partners. Importantly, this retention is more pronounced for genes susceptible to dominant deleterious mutations.

10.2 "Dangerous" genes show no biased retention by SSD or CNV

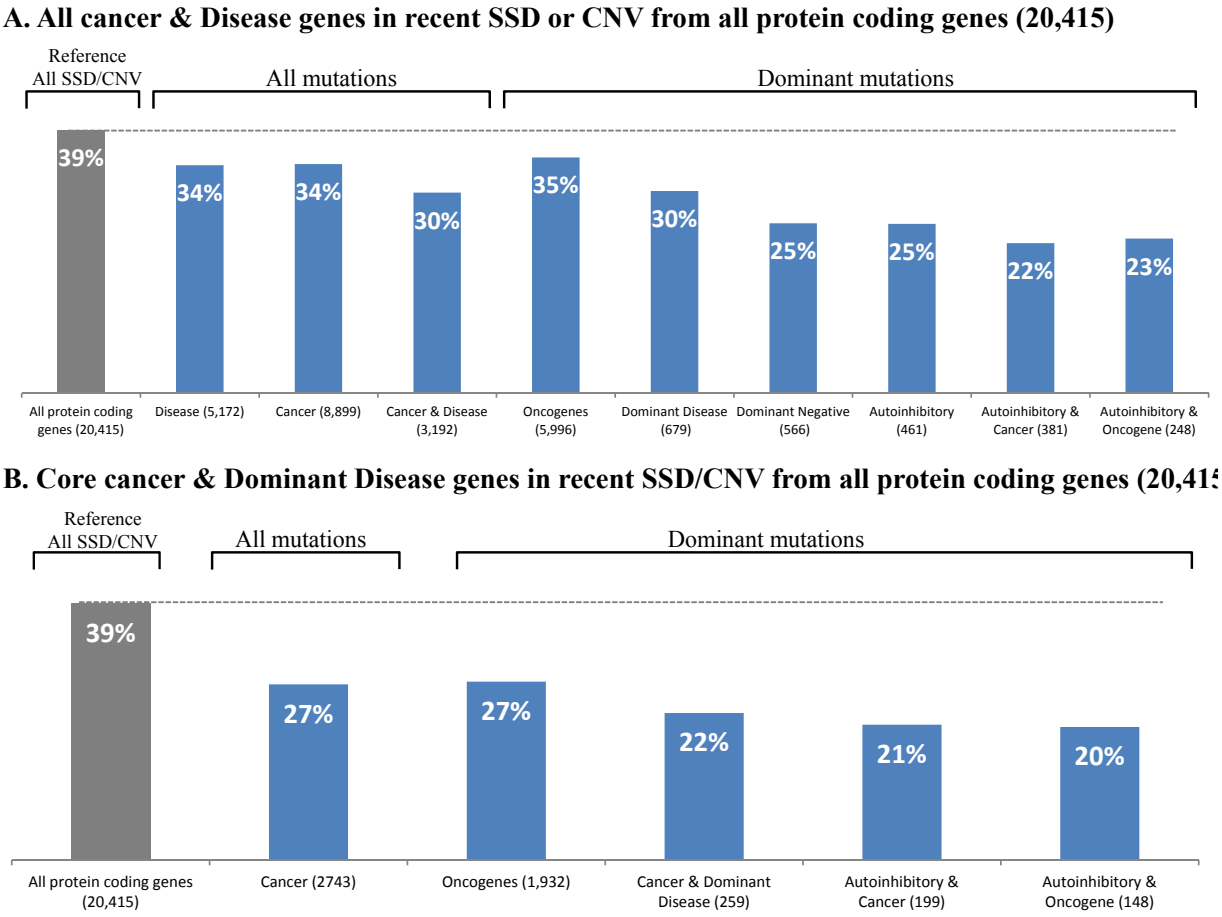

Figure 10.4: Retention of SSD/CNV duplicates in the gene classes prone to deleterious mutations

We then sought to investigate the retention of human disease genes by small scale duplicates (SSD). Unlike WGD, organisms frequently encounter SSDs in their genomes. The rate of SSD has been estimated to be 0.01 per gene per million years, which is comparable to single nucleotide mutations [Lynch and Conery, 2000, 2003]. Therefore, it is expected that majority of cancer genes must have also experienced SSDs during their evolutionary history. Copy number variations (CNVs) are another form of SSDs that are not yet fixed in the genome. To examine the proportion of "dangerous" genes retained by SSDs or that have variable copy numbers in the human genome, we reconciled our data in all the aforementioned disease gene classes with SSD or CNV duplicates from two different approaches.

10.2.1 Small scale duplicates from Ensembl

We first specifically focused on recent SSDs *i.e.* SSDs that occurred after the 2R-WGD as per relative duplication time estimations from Ensembl compara. We also merged data of recent SSDs (4,465) and CNV genes (5,185) leading to a total of 7978 locally duplicated genes. As depicted in Figure 10.4 A, human disease genes are significantly depleted in SSD/CNV genes. All cancer genes in the human genome have retained significantly less SSD duplicates, and they also remain refractory to CNVs in the human genome (34%; 3,032/8,899; p-value = 3.594×10^{-22}, χ^2 test), as opposed to the global proportion of 39% SSD/CNV duplicates.

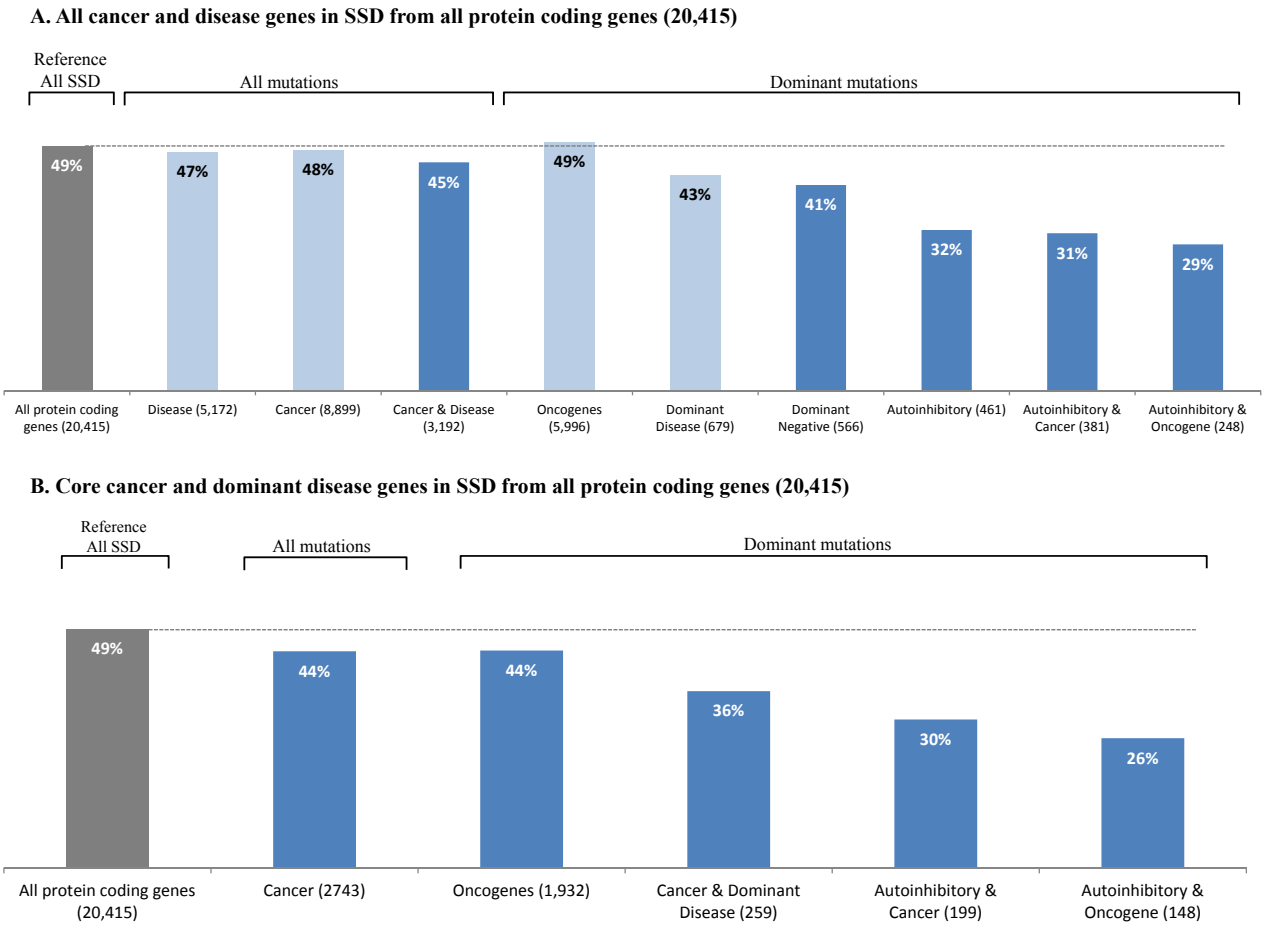

Figure 10.5: Retention of SSD duplicates from sequence comparison in the gene classes prone to deleterious mutations. Bars in light blue have non significant p-values.

Similarly, Mendelian disease genes have also retained fewer SSD/ CNV duplicates than expected by chance, (34%; 1,754/5,172; p-value = 2.132×10^{-14}, χ^2 test). We observed even weaker retention of SSD/CNV duplicates in gene classes prone to dominant deleterious mutations, with oncogenes (35%; 2,101/5,996; p-value = 1.455×10^{-10}, χ^2 test), dominant disease genes (30%; 204/679; p-value = 1.399×10^{-06}, χ^2 test), dominant negative genes (25%; 143/566; p-value = 1.632×10^{-11}, χ^2 test), and genes with autoinhibitory folds (25%; 116/461; p-value = 9.135×10^{-10}, χ^2 test), having significant depletion in SSD/CNV duplicates. Likewise, combining these properties also leads to even lower SSD/CNV retention biases, with cancer and Mendelian disease genes (30%; 952/3,192; p-value = 8.566×10^{-27}, χ^2 test), cancer and autoinhibitory genes (22%; 85/381; p-value = 1.966×10^{-11}, χ^2 test), and oncogenes with autoinhibitory domains (23%; 57/248; p-value = 2.049×10^{-07}, χ^2 test), retaining least SSD/CNV genes.

Consistent with above trends, we observed that genes belonging to our core cancer dataset also display significant reduction in paralogs formed by SSD or CNV Figure 10.4 B. All cancer genes from core dataset have retained only 27% SSD/CNV paralogs as opposed to 39% of the genome average (27%; 734/2,743; p-value = 6.356×10^{-40}, χ^2 test). Oncogenes from the core dataset also display similar depletion in SSD/CNV (27%; 525/1,932; p-value = 7.795×10^{-27}, χ^2 test), which becomes even stronger when the categories are combined together, e.g. cancer and dominant disease genes (22%; 58/259; p-value = 3.726×10^{-08}, χ^2 test), cancer and autoinhibitory genes (21%; 41/199; p-value = 9.206×10^{-08}, χ^2 test), and autoinhibitory oncogenes

(20%; 30/148; p-value = 2.737×10^{-06}, χ^2 test).

10.2.2 Small scale duplicates from sequence comparisons

Ensembl compara estimates duplication nodes by reconciling gene trees with the species tree of the organisms in Ensembl database[Vilella et al., 2009]. The estimation depends on the available species and the duplication node of a paralog pair can change from version to version. Therefore, we also identified 9,916 recent duplicates using sequence comparisons as described in the Section 6.11.

As shown in Figure 10.5, we again observed that either there is no bias for the disease genes' retention by SSD, or there is a substantial depletion, as seen for Ensembl SSD and CNV duplicates. Against a genome average of 49% SSDs, all disease genes (47%; 2,448/5,172; p-value = **0.08** test), all cancer genes (48%; 4,254/8,899; p-value = **0.14**χ^2 test), oncogenes (49%; 2,958/5,996; p-value = **0.28**χ^2 test) show no bias for SSDs. Dominant disease genes are only slightly depleted (43%; 291/679; p-value = **0.003**χ^2 test), whereas other classes that are highly susceptible to dominant deleterious mutations, such as dominant negative (41%; 231/566; p-value = 2.211×10^{-04}, χ^2 test), autoinhibitory genes (32%; 141/461; p-value = 7.625×10^{-13}, χ^2 test), autoinhibitory genes mutated in cancer (31%; 119/381; p-value = 1.275×10^{-11}, χ^2 test), or autoinhibitory oncogenes (29%; 72/248; p-value = 7.423×10^{-10}, χ^2 test), are highly depleted in SSD.

In conclusion, we observed that cancer and disease genes either show no biased retention of SSD, or they are strongly depleted (especially the gene classes very highly susceptible to dominant mutations) in both the dataset of recent SSD/CNV paralogs.

10.2.3 Ohnolog and SSD retention bias in different human primary tumors

We also investigated the ohnolog retention bias for different human primary tumor types from COSMIC database [Forbes et al., 2008]. Out of total 8,899 genes mutated in cancer 8,837 are part of COSMIC datbase. For each of the primary tumor types classified by COSMIC, we obtained the information of all the genes that are found to be mutated in them. Table 10.1 lists 20 tissues where the total mutated genes exceeds 100. Against the genomic reference of 36% ohnologs and 49% SSD duplicates (from sequence comparison), it can be clearly shown that all ohnologs are highly enriched in all the tumor types. χ^2 p-values range from 4.17×10^{-118} for lung cancer to the lowest, yet highly significant 1.99×10^{-05} for soft tissues.

In contrast, we observed that for all the cancer types, there is no significant bias in retention of SSD duplicates. Either the SSD retention in the human tumors is slightly less than the genomic average (e.g. for prostate cancer, χ^2 p-value 0.004) or does not deviate significantly for most of the other tissues. This clearly demonstrates that while cancer genes are highly likely to be retained after duplication by WGD, there is no bias for their retention by small scale duplications.

Table 10.1: Dangerous gene retention in ohnologs and SSD for different human primary tumors from COSMIC

Human Primary Tumor Types	**# of Genes**	**# of WGD**	**% of WGD**	**WGD P-value**	**# of SSD**	**% of SSD**	**SSD P-value**
All protein coding genes	20415	7351	36.0	*Reference*	9916	48.6	*Reference*
All primary tumors	8837	4222	47.8	1.58×10^{-117}	4203	47.6	0.06
Lung	8621	4134	48.0	4.17×10^{-118}	4108	47.7	0.09
Large intestine	8449	4059	48.0	1.75×10^{-117}	4014	47.5	0.05
Endometrium	7435	3620	48.7	7.49×10^{-115}	3537	47.6	0.08
Kidney	6218	3079	49.5	4.05×10^{-109}	2950	47.4	0.07
Ovary	5771	2889	50.1	1.42×10^{-109}	2733	47.4	0.06
Skin	5742	2826	49.2	1.50×10^{-96}	2753	47.9	0.34
Prostate	5428	2729	50.3	2.62×10^{-106}	2612	48.1	0.51
Breast	5143	2644	51.4	3.69×10^{-117}	2371	46.1	0.0004
Upper aerodigestive tract	3838	1962	51.1	1.01×10^{-84}	1897	49.4	0.29
Urinary tract	3312	1707	51.5	2.16×10^{-77}	1536	46.4	0.01
Central nervous system	3155	1683	53.3	1.72×10^{-91}	1463	46.4	0.01
Pancreas	3143	1579	50.2	4.97×10^{-62}	1472	46.8	0.05
Haematopoietic & lymphoid tissue	2612	1393	53.3	5.87×10^{-76}	1238	47.4	0.23
Stomach	1878	888	47.3	2.43×10^{-24}	941	50.1	0.18
Cervix	1810	912	50.4	3.37×10^{-37}	865	47.8	0.51
Liver	1179	623	52.8	2.16×10^{-33}	552	46.8	0.14
Oesophagus	917	490	53.4	4.09×10^{-28}	441	48.1	0.77
Autonomic ganglia	356	175	49.2	2.36×10^{-07}	187	52.5	0.11
Biliary tract	148	89	60.1	9.67×10^{-10}	61	41.2	0.07
Soft tissue	112	62	55.4	1.99×10^{-05}	41	36.6	0.15

10.3 Mapping cancer and disease gene duplications on Ensembl duplication nodes

To clearly emphasize the global picture of the retention of cancer and disease genes by the 2R-WGD in the human genome, Figure 10.6 A & B show the distribution of all the duplication events, involving at least one cancer and disease gene on the phylogenetic nodes leading to humans. Of the total 9,852 duplication events involving cancer genes, 56% can be traced back to the base of vertebrates, of which 73% belong to the 2R-WGD in our dataset. Similarly, for the disease genes, 59% of the 5,410 retained duplicates are mapped to the node Vertebrata, and 68% of these belong to 2R-WGD.

It can be clearly seen that after vertebrates, hardly any small scale duplicates of these "dangerous" genes have been retained. In fact, most of the duplication events not belonging to WGD can be traced back to nodes older than vertebrates, most notably coelomata. Although the possibility of additional WGD events before the 2R cannot be excluded, after more than 600-700 MY of evolution, traces of such duplication would be extremely difficult to prove.

We still suspect that some of the cancer and disease gene duplicates at the base of vertebrates that we could not identify as ohnologs (or were excluded by *P-value* filters) can still be "true" ohnologs. Nevertheless, the global picture of disease-ohnolog association is already highly significant.

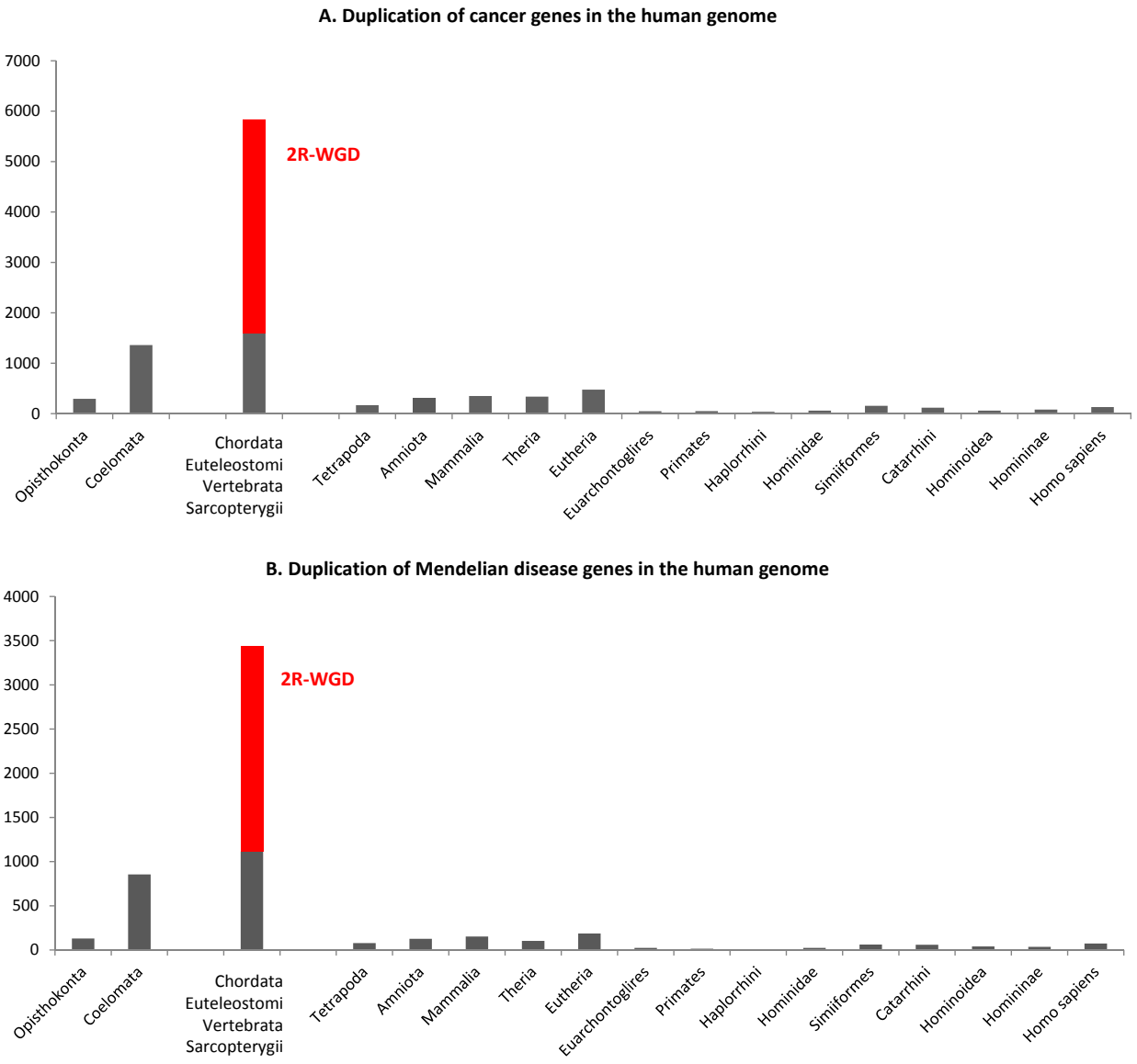

Figure 10.6: Duplication of A. cancer and B. Mendelian disease genes at each phylogenetic node leading to humans

10.4 Ohnologs are more conserved than non-ohnologs

We then investigated whether the susceptibility of ohnologs to deleterious mutations could be directly quantified through comparative sequence analysis. We used Ka/Ks ratio (also called D_N/D_S) estimates, which measure the proportion of nonsynonymous substitutions (Ka or D_N) to the proportion of synonymous substitutions (Ks or D_S). Ohnologs exhibit statistically lower Ka/Ks ratios (Table 10.2) which provides direct evidence of strong conservation, consistent with a higher susceptibility of ohnologs to deleterious mutations. Similar trends have also been reported for ohnologs specific to teleost fishes [Brunet et al., 2006] or to the more recent WGD in *Xenopus laevis* lineage [Sémon and Wolfe, 2008].

We used human orthologs from all the six invertebrate outgroups to calculate Ka/Ks ratios. As detailed in Table 10.2, non-ohnologs have significantly lower p-values of Mann–Whitney U (also called the Mann–Whitney–Wilcoxon or MWW) test. While for ohnologs from strict and

Table 10.2: Analysis of sequence conservation with respect to all invertebrate outgroups

	Human – Invertebrate pair	Mean	Median	n Distribution	Mean	Median	n Distribution	*p*-value MWW test
	Strict ohnologs *versus* Non-ohnologs							
		Strict ohnologs			Non-ohnologs			
A	*H. sapiens - B. floridae*	0.1283	0.1095	3,438	0.1755	0.1409	12,070	2.15×10^{-100}
B	*H. sapiens - C. intestinalis*	0.1708	0.1486	3,262	0.1915	0.1673	9,361	8.36×10^{-21}
C	*H. sapiens - C. savignyi*	0.1663	0.1415	3,199	0.1903	0.1633	9,118	6.75×10^{-27}
D	*H. sapiens - S. purpuratus*	0.1669	0.1445	3,007	0.1835	0.1585	8,787	2.13×10^{-12}
E	*H. sapiens - D. melanogaster*	0.1784	0.1632	3,282	0.1956	0.1818	9,066	5.89×10^{-26}
F	*H. sapiens - C. elegans*	0.2047	0.189	3,147	0.2223	0.205	8,511	1.39×10^{-19}
	Intermediate ohnologs *versus* Non-ohnologs							
		Intermediate Ohnologs			Non-Ohnologs			
A	*H. sapiens - B. floridae*	0.1355	0.1144	5,221	0.1801	0.1449	10,287	1.26×10^{-103}
B	*H. sapiens - C. intestinalis*	0.1745	0.1531	4,921	0.1936	0.1686	7,702	5.03×10^{-18}
C	*H. sapiens - C. savignyi*	0.1727	0.1479	4,857	0.1914	0.1643	7,460	3.03×10^{-20}
D	*H. sapiens - S. purpuratus*	0.1716	0.1487	4,479	0.184	0.159	7,315	4.52×10^{-08}
E	*H. sapiens - D. melanogaster*	0.1824	0.1675	4,907	0.1967	0.183	7,441	1.02×10^{-21}
F	*H. sapiens - C. elegans*	0.2081	0.1912	4,724	0.224	0.2066	6,934	2.45×10^{-20}
	Relaxed ohnologs *versus* Non-ohnologs							
		Relaxed Ohnologs			Non-ohnologs			
A	*H. sapiens - B. floridae*	0.144	0.122	6,697	0.181	0.1422	8,811	5.82×10^{-56}
B	*H. sapiens - C. intestinalis*	0.1833	0.1618	6,192	0.1889	0.1624	6,431	0.3682
C	*H. sapiens - C. savignyi*	0.1825	0.1559	6,118	0.1855	0.1579	6,199	0.0517
D	*H. sapiens - S. purpuratus*	0.1785	0.1568	5,647	0.1801	0.154	6,147	0.2838
E	*H. sapiens - D. melanogaster*	0.1892	0.1739	6,157	0.1928	0.1795	6,191	0.0024
F	*H. sapiens - C. elegans*	0.2152	0.1971	5,872	0.2199	0.2041	5,786	0.0013

intermediate probability filters, the Ka/Ks distributions are significantly lower for ohnologs than for non-ohnologs for all the six outgroups. For relaxed filters, we observed that this difference is reduced for *C. intestinalis* and Sea urchin, yet the Ka/Ks are still significantly lower for Amphioxus, Drosophila, Worm and *C. savignyi*.

The distributions of Ka/Ks ratios for ohnologs from all the three criteria, along with non ohnologs have been shown in Figure 10.7. All the three distributions are significantly different than the one for non-ohnologs from relaxed criteria (Table 10.2). We also observed that the Ka/Ks ratios from the strict criteria are also significantly different than the ones from intermediate criteria (Median = 0.109 for strict ohnologs *versus* 0.114 for intermediate ohnologs; p-value = 1.838×10^{-4}, MWW test). Similarly ohnologs from strict and relaxed criteria (Median = 0.109 for strict ohnologs *versus* 0.122 for relaxed ohnologs; p-value = 2.922×10^{-16}, MWW test), and intermediate and relaxed criteria also show significant differences in the distributions (Median = 0.114 for intermediate ohnologs *versus* 0.122 for relaxed ohnologs; p-value = 6.687×10^{-9}, MWW test). Note, however, that these p-values for MWW test are considerably lower than the difference of all the three criteria from non-ohnologs (Table 10.2 A). Yet, this difference signifies that ohnologs for lower *P-value* from our approach tend too be even more highly conserved.

Note, also, that the functional consequences of such deleterious mutations, leading either to a gain or a loss of function, cannot be directly inferred from these Ka/Ks distributions. Yet, as outlined in the next section, we found marked differences in the retention of “dangerous” ohnologs prone to dominant gain-of-function mutations and “essential” ohnologs exhibiting lethal loss-of-function or null mutations, or recessive disease mutations.

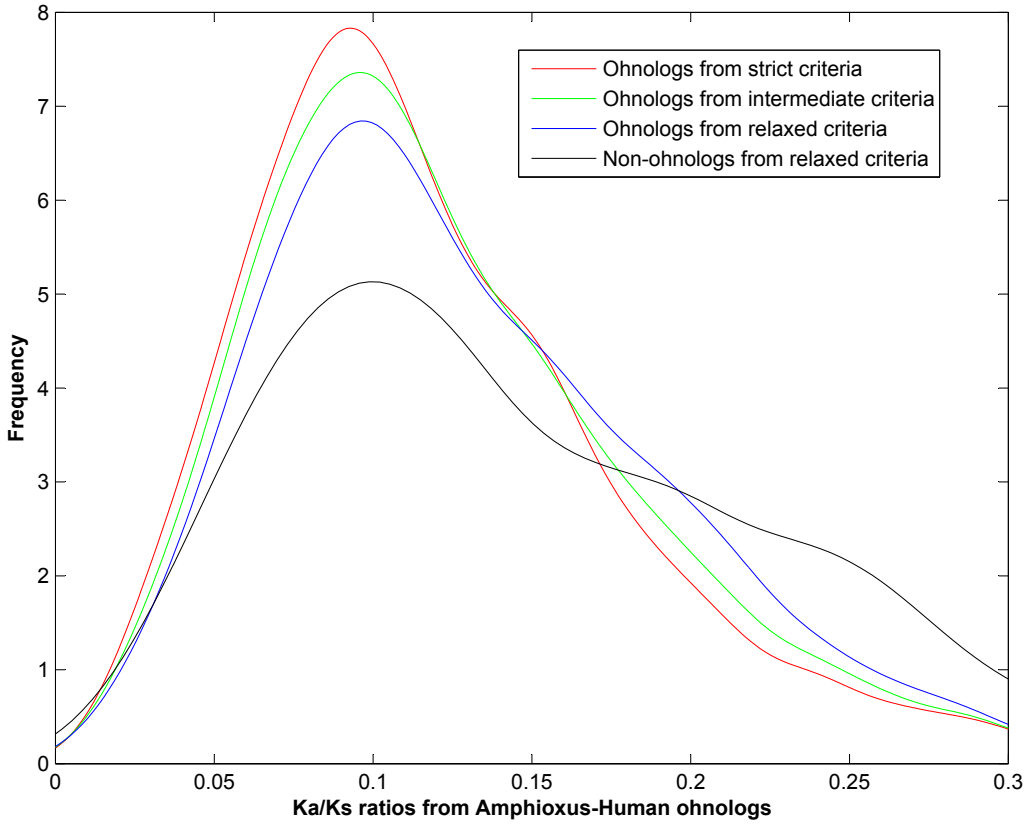

Figure 10.7: Ka/Ks distribution between human pairs identified to be ohnologs by strict, intermediate of relaxed P-values, or non-ohnologs. *X-axis*, Ka/Ks ratios; *Y-axis*, density.

10.5 Dominant, and not recessive disease genes have retained more ohnologs

As noted in Section 10.1 above, the ohnolog retention is most enhanced for the genes susceptible to dominant deleterious mutations such as oncogenes, autoinhibitory or dominant negative genes. To further characterize the ohnologs with respect to the inheritance pattern of mutations, we explored the retention bias of genes susceptible to recessive disease mutations.

10.5.1 Recessive disease genes

First, we focused on recessive disease genes in our dataset. We observed that, unlike dominant disease genes, recessive disease genes do not present any significant retention bias from 2R-WGD. 888 experimentally verified autosomal recessive disease genes from OMIM have not retained more ohnologs than expected by chance in the human genome, 39% *vs* 35% (35%; 314/888; p-value = **0.68**, χ^2 test). The observation is also consistent on a slightly larger dataset of 1,002 genes (37%; 367/1,002; p-value = **0.67**, χ^2 test), where we just excluded the genes known to be implicated in dominant disorders, but kept all other inheritance patterns provided by OMIM (e.g. polygenic, multifactorial, heterogeneous or uncertain).

10.5.2 Essential genes

We then, estimated the retention pattern of "essential" genes in ohnologs. Essential genes were defined as the genes whose knock-out leads to lethality of infertility. Unless haploinsufficient, such mutations are typically recessive. As described in Section 7.6, we have obtained essential genes from two different resources, from OGEE [Chen et al., 2012; Silva et al., 2008] based on knock-out experiments in human cell lines, and orthologs of mouse essential genes from MGI [Eppig et al., 2012], based large-scale null mutant studies in mouse.

Of the 6,436 orthologs of mouse genes tested for essentiality, 46% (2,938) have been identified to be essential. The global proportion of ohnologs in these 6,436 genes is 52%(3,319). We observed that the human orthologs of the mouse essential genes, are only slightly enriched in ohnologs 54% *vs* 52% (54%; 1,597/2,938; p-value = $\mathbf{0.008}, \chi^2$ test), where 54% is the proportion of ohnologs among the 2,938 genes found to be essential in mouse. Furthermore, this small enrichment becomes hardly significant, once genes with dominant allelic mutants are removed from the list of 6,436 genes tested for essentiality in mouse 47% *vs* 44% (47%; 664/1,406; p-value = $\mathbf{0.03}, \chi^2$ test), where 44% = 1,519/3,427 is the global proportion of ohnologs among the 3,427 genes tested for essentiality in mouse, after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or auto-inhibitory folds.

These results are also consistent for ohnologs from intermediate and strict criteria. When we explored the enrichment of the essential genes in 5,733 ohnologs from intermediate criteria, we could find the low enrichment of essential genes in ohnologs, 44% (44%; 1,306/2,938; p-value = $5.527 \times 10^{-04}, \chi^2$ test) against 41% intermediate ohnologs in all genes tested for essentiality. For the 3,783 ohnologs from the strict *P-value* criteria, which we expect to be very strong ohnolog candidates, essential genes display even lower enrichment with only 32% high confidence ohnologs constituting essential genes (32%; 928/2,938; p-value = $\mathbf{0.004}, \chi^2$ test), against a global retention of 29% (1,843/6,436) ohnologs in all genes tested. For both these cases, after removal of "dangerous" ohnologs, this slight prevalence becomes non-significant. After removal of "dangeous" genes, the proportion of ohnologs from intermediate criteria is 38% (38%; 530/1,406; p-value = $\mathbf{0.02}, \chi^2$ test), against a global retention of 35% intermediate ohnologs in all tested genes. Similarly, only 25% ohnologs from strict criteria are essential (25%; 354/1,406; p-value = $\mathbf{0.06}, \chi^2$ test) against global average of 23% strict ohnologs in all tested genes after removal of dominant disease genes.

We also explored the ohnolog retention statistics on the 18,938 genes tested for essentiality, directly in human cell lines. We noticed that essential genes from these cell line knock-outs are significantly enriched in the ohnologs form relaxed criteria, 48% *vs* 37% (48%; 1,280/2,701; p-value = $4.157 \times 10^{-27}, \chi^2$ test), where 37% = 7,074/18,938 is the global proportion of ohnologs in the genes tested for essentiality. This enrichment also persists, when we restrict ourselves to genes not implicated in dominant diseases as discussed above. Ohnologs without dominant diseases are also enriched in essential genes from the cell line knock outs, however, the bias is less than that for all genes, 39% *vs* global ohnolog retention of 31% (3,793) in 12,283 genes tested for essentiality after removal of "dangerous" genes, (39%; 519/1,344; p-value = $8.321 \times 10^{-10}, \chi^2$ test).

We then looked for this enrichment in the ohnologs from intermediate and strict *P-value* criteria. Unlike orthologs of mouse essential genes, we still could find that essential genes from OGEE are enriched in both intermediate (37%; 1,004/2,701; p-value = $2.749 \times 10^{-20}, \chi^2$ test), and strict ohnologs (26%; 711/2,701; p-value = $6.431 \times 10^{-09}, \chi^2$ test), against a global retention of 29% (5,512/18,938) intermediate and 19% (3,645/18,938) strict ohnologs in essential genes. This enrichment persisted even if we removed the genes susceptible to dominant deleterious mutations, however, it was reduced considerably. In 12,283 genes with essential-

ity information and without dominant disease genes, the global proportion of ohnologs is 23% and 15% respectively from intermediate and strict *P-value* filters. Yet, 1,344 essential genes have retained 29% (29%; 393/1,344; p-value = $6.431 \times 10^{-20}, \chi^2$ test), and 21% (21%; 285/1,344; p-value = $1.976 \times 10^{-09}, \chi^2$ test), of their ohnolog partners from intermediate and strict criteria respectively.

Finally, we looked at the enrichment of mouse essential genes in mouse ohnologs directly. We observed that 3,242 (50%) of the total 6,435 gene tested for essentiality in mouse were ohnologs. The global proportion of mouse ohnologs from strict criteria in essential genes however was only 53% (53%; 1,548/5,268; p-value = $\mathbf{0.012}, \chi^2$ test). Furthermore, mouse ohnologs from intermediate and strict criteria are only slightly enriched in essential genes. Against the global retention of 42% mouse ohnologs from intermediate criteria in all tested genes, 46% esential genes have retained ohnologs (46%; 1,343/2,938; p-value = $\mathbf{0.0002}, \chi^2$ test). Against the 25% strict ohnologs in the genes tested for essentiality, proportion of ohnologs in essentila genes is 29% (29%; 830/2,938; p-value = $7.42 \times 10^{-5}, \chi^2$ test).

To study if this slight enrichment is contributed in fact by disease genes, we removed 1,242 mouse disease genes from our dataset. After removing the mouse disease genes, we observed that this slight enrichment of ohnologs in essential genes diminishes considerably for all the three datasets. Against the global proportion of 50%, 42% and 25% ohnologs in the 5,438 tested genes (after removal of disease genes) from relaxed, intermediate and strict *P-value* criteria respectively, we observed that ohnologs show very weak enrichment: (52%; 1,177/2,270; p-value = $\mathbf{0.03}, \chi^2$ test) for relaxed, (45%; 1,020/2,270; p-value = $\mathbf{0.001}, \chi^2$ test) for intermediate and (28%; 640/2,270; p-value = $\mathbf{0.002}, \chi^2$ test) for strict ohnologs.

In conclusion therefore, we observed that there is no robust enrichment of essential genes in ohnologs in the human or mouse genome. The results from human cell lines also show considerable reduction after removal of "dangerous" genes. We argue that the results for essential genes from cell lines knock outs should be interpreted with caution, as the genes not found to be essential in mammary cells can be essential for other cell lines and *vice-versa*. In addition, a gene found to be essential for survival of individual cells may not always have adverse effect on the fitness of organism, owing to the removal of slow dividing cells by cell-cell competition [Baker, 2011].

All in all, this shows that the retention of ohnologs has been most enhanced for genes prone to autosomal-dominant deleterious mutations and not autosomal-recessive deleterious mutations. This suggests that the retention of ohnologs is more strongly related to their "dangerousness," as defined by their high susceptibility to dominant deleterious mutations, than their functional importance or "essentiality". Ultimately, we will argue that the "dangerousness" of ohnologs effectively controls their individual retention in the genomes of post-WGD species, as will be shown below in the section Model for the Retention of Dangerous Ohnologs.

11

Dosage Balance, Expression level & Human Ohnologs

We have seen in the previous chapter that the susceptibility to dominant deleterious mutation is highly correlated with the retention of ohnologs, but not SSD in the vertebrate genomes, suggesting that gene families prone to dominant deleterious mutations have been particularly expanded by WGD. Antagonistic retentions of ohnologs and SSD across different functional categories has also been observed previously [Davis and Petrov, 2005; Hakes et al., 2007]. Yet, many of the evolutionary concepts explaining the retention of genes after duplication (e.g. neofunctionalization, subfunctionalization) were developed before the realization of this antagonistic retention pattern of ohnologs and SSD, and the overall impact of polyploidy on genome evolution.

However, an alternative hypothesis, focusing on the collective retention of interacting ohnologs, has been frequently invoked to account for the biased retention of ohnologs in unicellular organisms like yeast [Papp et al., 2003] or the paramecium [Aury et al., 2006] and in higher eukaryotes [Birchler et al., 2001; Makino and McLysaght, 2010]. This "dosage balance" hypothesis posits that interacting protein partners tend to maintain balanced expression levels in the course of evolution, in particular, for protein subunits of conserved complexes [Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Makino and McLysaght, 2010]. Thus, SSD of dosage balanced genes are thought to be generally detrimental through the dosage imbalance they induce, thereby raising the odds for their rapid nonfunctionalization [Papp et al., 2003; Maere et al., 2005]. By contrast, rapid nonfunctionalization of ohnologs after WGD has been suggested to be opposed by dosage effect, in particular, for highly expressed genes, and genes involved in protein complexes or metabolic pathways [Aury et al., 2006; Evlampiev and Isambert, 2007; Gout et al., 2010; Makino and McLysaght, 2010]. This is because WGD initially preserves correct relative dosage between expressed genes, whereas subsequent random nonfunctionalization of individual ohnologs disrupts this initial dosage balance.

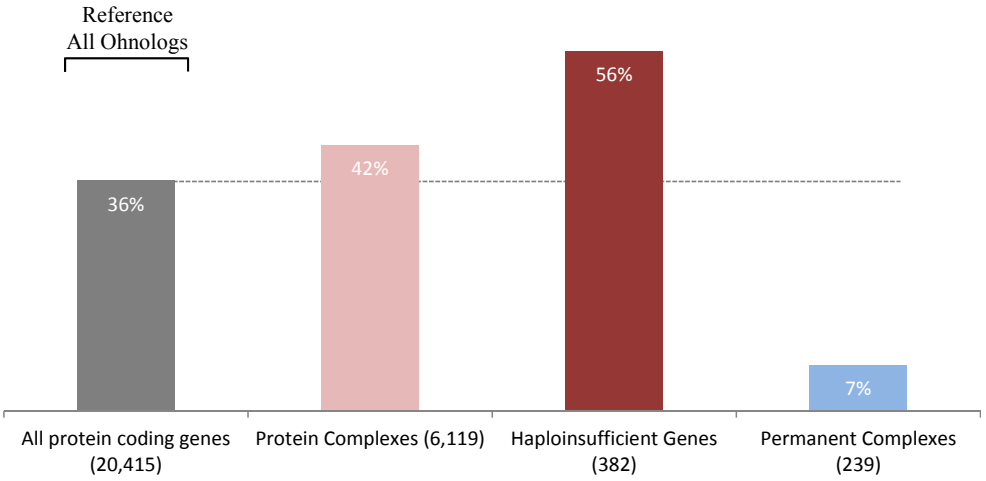

Figure 11.1: Enrichment of dosage balanced genes in the human ohnologs from relaxed criteria. Total number of the genes in the genome are given in brackets and the bars represent the fraction in the human ohnolgs

11.1 Mixed susceptibility of human ohnologs to dosage balance

11.1.1 High retention of protein complexes in ohnologs

The dosage balance hypothesis has been supported by the enrichment of protein complex subunits in ohnologs in a variety of organisms [Papp et al., 2003; Aury et al., 2006; Makino and McLysaght, 2010]. Following these results, we studied the enrichment of human protein complexes in ohnologs. To this end, we compiled a dataset of genes involved in the formation of human protein complexes from multiple resources (Section 7.5). A total of 6,119 genes could be identified to be associated to experimentally verified and manually curated protein complex subunits in the human genome.

As depicted in Figure 11.1, we observed in agreement with [Makino and McLysaght, 2010], that the genes implicated in multiprotein complexes have retained significantly more ohnologs from relaxed criteria than expected by chance, 42% *versus* 36% (42%; 2,541/6,119; p-value = 2.406×10^{-19}, χ^2 test). This trend is also enhanced when focusing on haploinsufficient genes, that are known for their actual sensitivity to dosage balance constraints [Qian and Zhang, 2008] (56%; 215/382; p-value = 1.516×10^{-16}, χ^2 test). Haploinsufficient mutations are in fact dominant disease mutations, affecting normal phenotype by gene dosage. Therefore, their enrichment in ohnologs is also expected from their susceptibility to dominant deleterious mutations as discussed in Section 10.5.

These observations are also consistent for the ohnologs from intermediate and strict *P-values*. Against genome average of 28% ohnologs from intermediate criteria, all protein complex genes (34%; 2,072/6,119; p-value = 3.065×10^{-22}, χ^2 test), and haploinsufficient genes (46%; 174/342; p-value = 6.596×10^{-14}, χ^2 test), have retained more ohnologs. Similarly, ohnologs from the strict criteria also display enhanced ohnolog retention in all complexes (24%; 1,442/6,119; p-value = 3.760×10^{-24}, χ^2 test), and haploinsufficient genes (36%; 136/382; p-value = 8.891×10^{-18}, χ^2 test), against the genome average of 19%.

11.1.2 Transient *versus* permanent complexes

We then focussed our attention on manually curated dataset of the 239 genes involved in formation of permanent complexes. While most of the complex 6,119 genes involved in formation of transient complexes show an enrichment in ohnologs. Surprisingly, an opposite trend corresponding to the elimination of ohnologs is observed for genes implicated in permanent complexes, that are presumably strongly sensitive to dosage balance constraints. (7%; 17/239; p-value = 1.329×10^{-20}, χ^2 test).

These observations are also consistent for intermediate and strict ohnologs. A strong depletion for manually curated permanent complexes (7%; 16/239; p-value = 1.270×10^{-13}, χ^2 test) for 28% ohnologs from intermediate criteria was observed. Likewise, even from strict *P-value* ohnologs, a mere (5%; 11/239; p-value = 2.995×10^{-8}, χ^2 test) ohnologs have been retained in permanent complexes, against 19% of genome average. A recent study has also observed on a larger dataset of 80,202 interactions between 12,839 gene products for various protein-protein interaction databases that ohnologs do not have greater propensity to be retained in the same protein complex than expected by chance [Chen et al., 2013b].

While in the human genome most of the ribosomal genes have lost their ohnologs, this trend is opposite to what has been observed for *Saccharomyces cerevisiae* ohnologs where 76% of its ribosomal gene ohnologs from a 150 MY old WGD [Kellis et al., 2004; Lin et al., 2007] have been retained, although the maintenance of these ohnologs has been suggested to require frequent gene conversion events [Kellis et al., 2004; Evangelisti and Conant, 2010] as well as fine-tuned dosage compensation to ensure a balanced expression with the remaining 24% ribosomal genes having lost their ohnologs [Zeevi et al., 2011].

In fact, looking more closely at the few human ohnologs, that have not been eliminated from permanent complexes Table 11.1, we found that they are likely under less stringent dosage balance constraints than most proteins in permanent complexes, as they typically coassociate with mitochondrial proteins or form large multimeric subcomplexes with intrinsic stoichiometry disequilibrium.

This suggests that the elimination of most ohnologs from permanent complexes is, in fact, strongly favored under dosage imbalance and becomes likely inevitable once a few of those ohnologs have been accidentally lost following WGD. Indeed, the uneven elimination of ohnologs in permanent complexes is expected to lead to the assembly of nonfunctional, partially formed complexes detrimental to the cell, unless dosage compensation mechanisms effectively re-establish proper dosage balance at the level of gene regulation [Birchler et al., 2001], as for yeast ribosomal proteins [Zeevi et al., 2011]. By contrast, transient complexes, which are typ-

Table 11.1: **Low retention of ohnologs from the strict criteria in permanent complexes**

Permanent complexes	Ohnologs	Intrinsic stoichiometry disequilibrium of ohnologs in permanent complexes
ATP F0	3 / 12	The 3 ohnologs ATP5G1-3 form the 10-mer C-ring of the F-type ATP synthase
ATP F1	0 / 5	
COX	2 / 11	The 2 ohnologs COX4I1,2 coassemble with 3 mitochondrial encoded genes
SRS	1 / 32	RPS27L has an ohnolog RPS27, both are part of 40s subunit
Mitochondrial SRS	0 / 30	
LRS	2 / 50	RPL3 and RPL22 have ohnologs RPL3L and RPL22L1 with unknown functions
Mitochondrial LRS	0 / 48	
Proteasome	1 / 31	Ohnologs PSMA7 or PSMA8 are included in the 2 rings of 7 alpha subunits
PD	2 / 5	PDK1/2 are part of mitochondrial PD
RNA Pol II	0 / 12	
RNA Pol III	0 / 9	
RNA Pol III	0 / 9	

COX, Cytochrome-c Oxidase; LRS, Large Ribosomal Subunit; PD, Puryvate Dehydrogenase; SRS, Small Ribosomal Subunit.

ically more modular than permanent complexes, are expected to accommodate such dosage changes more easily, as they do not usually require the same strict balance in the expression levels of their protein partners.

These findings on the differences in retention of human ohnologs between permanent and more transient complexes suggest the relevance of different underlying causes. Although dosage balance presumably remains the primary evolutionary constraint in permanent complexes (<2% of human genes), which lead to the elimination of ohnologs in permanent complexes in vertebrate genomes, gene susceptibility to deleterious mutations may be more relevant for the retention of ohnologs within the 17% of human genes participating in more transient complexes. For instance, transient complexes involved in phosphorylation cascades or GTPase signaling pathways are known to be more sensitive to the level of activation of their protein partners than to their total expression levels. Thus, although the active forms of multistate proteins typically amount to a small fraction of their total expression level, hence providing a large dynamic range for signal transduction, it also makes them particularly susceptible to gain-of-function mutations. Such mutations can shift protein activation levels 10- to 100-fold without changes in expression levels and likely underlie stronger evolutionary constraints than the 2-fold dosage imbalance caused by gene duplication.

11.1.3 Susceptibility of human protein complexes to disease mutations

To further investigate the relative effects of dosage balance and gene susceptibility to deleterious mutations, we analyzed whether the overall enhanced retention of ohnologs within multiprotein complexes Figure 11.1 could indirectly result from an enhanced susceptibility to deleterious mutations.

Indeed, as outlined in Figure 11.2, human disease genes are more prevalent in protein complexes than expected by chance. While, an average of 30% genes are associated with the formation of protein complexes, they show enhanced association with cancer (39%; 3,441/8,899; p-value = 1.137×10^{-71}, χ^2 test), Mendelian diseases (42%; 2,188/5,172; p-value = 1.757×10^{-83}, χ^2 test), and oncogenes (34%; 2,020/5,996; p-value = 3.368×10^{-10}, χ^2 test). A slightly greater bias towards protein complexes is observed for both cancer & disease genes (48%; 1,522/3,192; p-value = 1.004×10^{-105}, χ^2 test), autoinhibitory cancer genes (45%; 170/381; p-value = 4.372×10^{-10}, χ^2 test), and autoinhibitory oncogenes (47%; 116/248; p-value = 7.687×10^{-09}, χ^2 test). Dominant disease genes (55%; 373/679; p-value = 9.578×10^{-46}, χ^2 test) dominant negative genes (64%; 365/566; p-value = 7.790×10^{-72}, χ^2 test) and genes with autoinhibitory protein fold (63%; 291/461; p-value = 1.975×10^{-54}, χ^2 test) are highly biased towards complexes. By contrast, ohnologs are only slightly, although significantly, more prevalent in complexes than expected by chance, 35% *versus* 30% (35%; 2,541/7,351; p-value = 5.102×10^{-79}, χ^2 test). In section Section 12.1, we will use a more advance statistical analysis to integrate this multivariate correlations between ohnolog retention, and their association to protein complexes and genetic disorders.

11.2 Gene expression level and human ohnologs

Gene expression levels have also been proposed to be correlated with the retention of ohnologs in yeast [Seoighe and Wolfe, 1999] and paramecium [Gout et al., 2009, 2010] genomes. In both these cases, it has been observed that genes with relatively high expression levels retain more

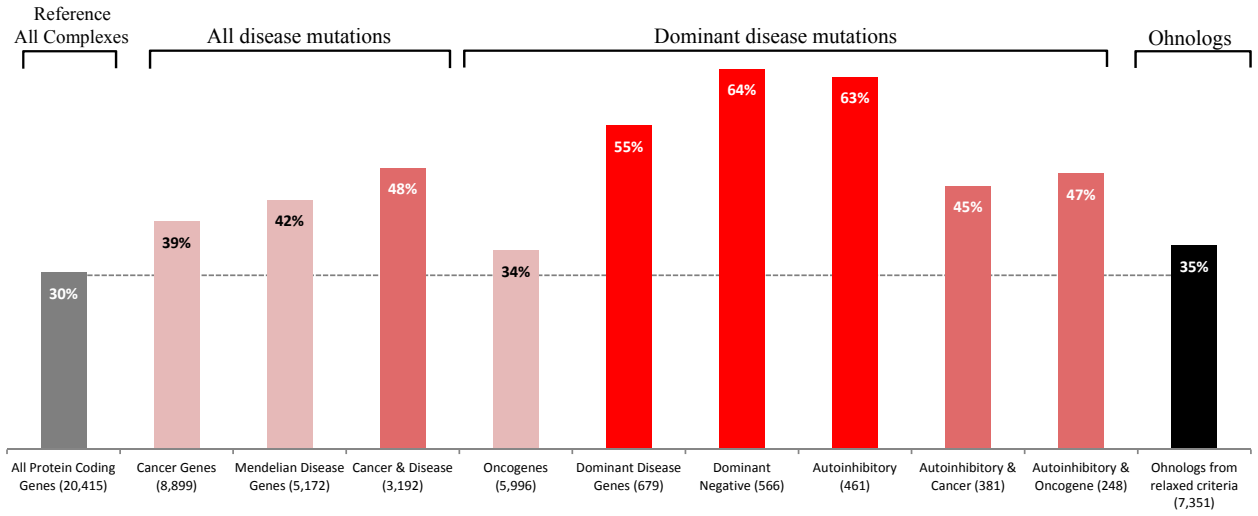

Figure 11.2: Enrichment of genes associated to protein complexes in human disease genes. The total number of genes in each category in the genome are given in brackets. Each bar represents the fraction of genes associated to complexes in each category.

ohnologs. Therefore, we also evaluated if high expression levels is also correlated with the ohnolog retention in the human genome.

We obtained expression levels of 13,425 human genes from BioGPS [Wu et al., 2009] as described in Section 7.8. The global proportion of relaxed ohnologs in these genes was 40%. If the higher expression levels affect ohnolog retention, we would expect that highly expressed genes would retain more ohnologs than expected by chance *i.e.* > 40%. We used three different percentile ranks of expression levels from our dataset to define highly expressed genes: 90^{th}, 75^{th} and median, corresponding to 61.9, 16.2 and 6.9 average difference (AD) values respectively. For each category, genes having expression values greater than these were considered to be highly expressed genes, leading to 1,342 genes corresponding to the highest 10% AD values, and 3,355 and 6,695 genes belonging to the top 25% and 50% expression levels respectively.

Contrary to the observations for yeast and paramecium, we observed no significant difference in the retention of human ohnologs between highly expressed genes and genome average for all the three categories (Figure 11.3). Genes belonging to the highest 10% expression levels were in fact slightly depleted in the ohnologs from relaxed criteria, 35% *versus* 40%, (35%; 471/1,342; p-value = **0.0002**, χ^2 test). Similarly, genes having top 25% expression values (39%; 1,293/3,355; p-value = **0.07**, χ^2 test), and expression levels greater than median (40%; 2,685/6,695; p-value = **0.97**, χ^2 test), have not retained more ohnologs than expected by chance.

These observations are also robust for the high confidence ohnologs from intermediate criteria. Highly expressed genes having top 10% (28%; 375/1,342; p-value = **0.005**, χ^2 test), top 25% (31%; 1,032/3,355; p-value = **0.33**, χ^2 test), and top 50% (32%; 2,144/6,695; p-value = **0.39**, χ^2 test) expression levels have not retained more ohnologs than the genome average of 32% from the intermediate *P-value* criterion. Similarly, against a genome average of 21% strict ohnologs in all genes with expression levels, there is no significant difference in the ohnologs retained from any of the three expression categories: highest expression level (18%; 245/1,342; p-value = **0.014**, χ^2 test), top 25% expression levels (21%; 711/3,355; p-value = **0.74**, χ^2 test), or expression levels greater than median value (22%; 1,476/6,695; p-value = **0.02**, χ^2 test).

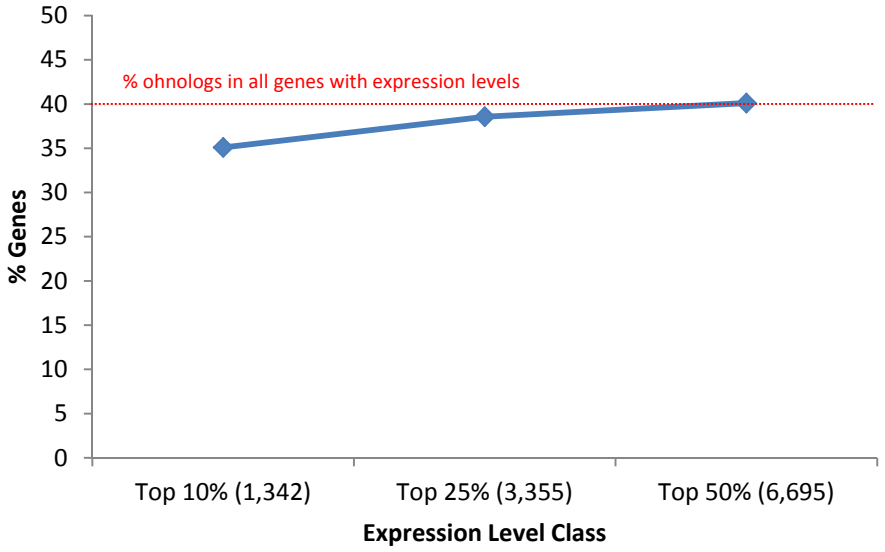

Figure 11.3: Percentage ohnologs in highly expressed genes in the human genome using three percentile ranks to define highly expressed genes: 90^{th}, 75^{th} and median.

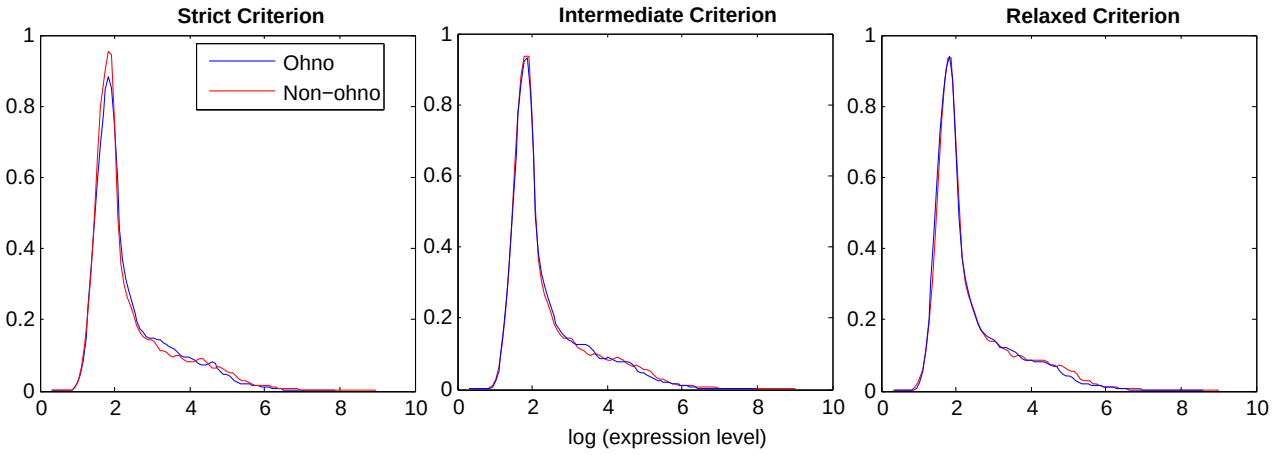

Figure 11.4: Expression level distribution for ohnologs and non-ohnologs for all the three *P-value* criteria

We next compared the distribution of expression levels of all ohnologs and non-ohnologs, and discovered the same trend. The distribution of the expression levels of all ohnologs and non-ohnologs from the three *P-value* criteria are shown in Figure 11.4. We could not see any significant difference in the distribution of expression levels between ohnologs and non ohnologs from strict (2,816 ohno *versus* 10,611 non-ohno; p-value = **0.11**, MWW test), intermediate (4,234 ohno *versus* 9,191 non-ohno; p-value = **0.88**, MWW test), and relaxed criteria (5,387 ohno *versus* 8,038 non-ohno; p-value = **0.025**, MWW test).

11.3 Sequence conservation and ohnolog retention

In addition to the aforementioned attributes, a number of previous studies have associated slow evolutionary rate with high duplicability and higher retention of ohnologs after WGD [Davis and Petrov, 2004; Brunet et al., 2006; Sémon and Wolfe, 2008; Jiang et al., 2013]. We also

analyzed Ka/Ks for the human ohnologs and observed that they tend to be highly conserved with lower Ka/Ks ratios, as discussed in detail in (Section 10.4) [Singh et al., 2012].

All in all, we observed a mixed susceptibility of human ohnologs for dosage balance, with all protein complexes retaining more ohnologs, while permanent complexes have clearly lost most of their ohnolog gene copies. Ohnologs also tend to have lower Ka/Ks ratios. However, we could not observe a significant bias towards retaining highly expressed ohnologs. In the next chapter, I will discuss in details the relative contribution of each of these properties in the retention of ohnologs using Mediation analysis.

12

Indirect Causes of Ohnolog Retention

In the previous chapters, we observed that the genes susceptible to dominant deleterious mutations have retained more ohnologs, clearly suggesting that the susceptibility to dominant deleterious mutations is a critical factor in the retention of ohnologs after WGD. We further observed, in agreement with previous studies [Papp et al., 2003; Makino and McLysaght, 2010], that protein complex genes that are sensitive to dosage balance constraints have also retained more ohnologs. However, many genes susceptible to dosage imbalance also display elevated mutation susceptibility, reflected in their enrichment in the disease genes classes in the human genome (Figure 11.2).

Evolutionary constraint to maintain balance in the relative dosage of the interacting subunits in macromolecular complexes has been frequently invoked to explain the differential retention of genes after SSD and WGD. However, based on our observations, we hypothesized that the susceptibility of individual genes to deleterious mutations directly underlies their retention after WGD. To further investigate the relative effects of dosage balance and gene susceptibility to deleterious mutations beyond these simple statistical associations, we quantified the total, direct and indirect effects of deleterious mutations and dosage balance constraints on the biased retention of human ohnologs.

To this end, we performed a Mediation analysis following the approach of Judea Pearl [Pearl, 2001, 2012a]. The Mediation framework, developed in the context of causal inference analysis, enabled us to evaluate the strength of total causal effect of these genomic properties on ohnolog retention. More importantly, it allowed us to assess the importance of mediators in transmitting the indirect effect of any property on the outcome of gene duplication via 2R-WGD.

Mediation analyses have been typically used in social sciences research [Baron and Kenny, 1986] as, for instance, in the context of legal disputes over alleged discriminatory hiring. In such cases, the problem is to establish that gender or race (X) have directly influenced hiring (Y) and not simply indirectly through differences in qualification or experience (M). Mediation analyses have also been used in epidemiology, as in a formal study [Robins and Greenland, 1992] that establishes the direct effect of smoking (X) on the incidence of cardiovascular diseases (Y), while taking into account the indirect effect of other aggravating factors, such as hyperlipidemia (M).

In this report, we have applied the Mediation analysis to genomic data (Chapter 8) to discriminate between direct effect (DE) and indirect effect (IE) of deleterious mutations (X or M) and dosage balance constraints (M or X) on the biased retention of human ohnologs (Y).

12.1 The effect of dosage balance is mediated by mutation susceptibility

First of all, we evaluated the effect of dosage balance constraints and deleterious mutation susceptibility on the retention of human ohnologs. To this purpose, we performed Mediation analysis with the three binary variables: dosage balance ('Dosage.Bal.'), deleterious mutation susceptibility ('Delet.Mut.') and 'Ohnologs'.

Gene classes susceptible to deleterious mutations include cancer, Mendelian disease, dominant negative and autoinhibitory genes; while protein complexes and haploinsufficient genes constitute the dosage balanced genes. A detailed analysis of the association tables and four mediation models involving ohnologs from the relaxed criteria, with all dosage balanced and disease genes is presented in the following sections (Sections 12.1.1, 12.1.2, 12.1.3 and 12.1.4). The results of these models, summarized in Figure 12.1 and Table 12.1, demonstrate that the retention of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multi-protein complexes.

Indeed, using the largest gene set, the direct causal effect of a change from "non-dosage balance" to "dosage balance" proteins only accounts for 58.4% of a small total effect (TE) of dosage balance on the retention of ohnologs ($DE/TE = 58.4\%$ with $TE = 0.086$), whereas 51% of this small total effect is indirectly mediated by their susceptibility to deleterious mutations ($IE/TE = 50.7\%$ with 9.1% non-linear coupling between direct and indirect effects, Section 12.1.1). By contrast, the alternative hypothesis, assuming a direct effect of deleterious mutations, accounts for 97.1% of a more than twice as large total effect on ohnolog retention ($DE/TE = 97.1\%$ with $TE = 0.200$), whereas the "dosage balance" *versus* "non-dosage balance" status of human genes has a negligible indirect effect on ohnolog retention in this causal relation hypothesis ($IE/TE = 6.1\%$, Section 12.1.2).

These trends are also further enhanced when the analysis is restricted to the 61% of human genes (12,437) without SSD and CNV duplicates. In fact, for the human genes without SSD and CNV, the total effect of dosage balance is even smaller, where most of this effect is indirectly mediated by deleterious mutation susceptibility ($IE/TE = 87.1\%$, with $TE = 0.056$). The direct effect of protein complex membership and haploinsufficiency also becomes lower ($DE/TE = 32.3\%$, Section 12.1.3). By contrast, the trend for the total effect of disease genes on ohnolog retention is further strengthened with $TE = 0.217$, and mostly transmitted *via* the direct route ($DE/TE = 100.7\%$), with negligible indirect mediation ($IE/TE = 3.9\%$, Section 12.1.4)

In fact, haploinsufficient genes can also be considered as susceptible to deleterious mutations. Hence, to test the robustness of our observations, we then studied the direct effect of protein complex subunits (removing dominant haploinsufficient genes) and its indirect mediation by cancer genes on the retention of ohnologs (Table 12.1A). Consistently, we observed that the total effect of protein complexes on ohnolog retention is very low ($TE = 0.078$) and a large proportion of this effect is again mediated by cancer associated genes ($IE/TE = 47.7\%$). The alternative model to assuming a direct effect of cancer genes on the retention of ohnologs is also in agreement with the larger dataset, with a very high total and direct effect, and negligible mediation by protein complexes genes ($TE = 0.205, DE/TE = 97.9\%, IE/TE = 4.4\%$, 12.1B). Removing SSD and CNV genes from these models also confirms a very small total effect of protein

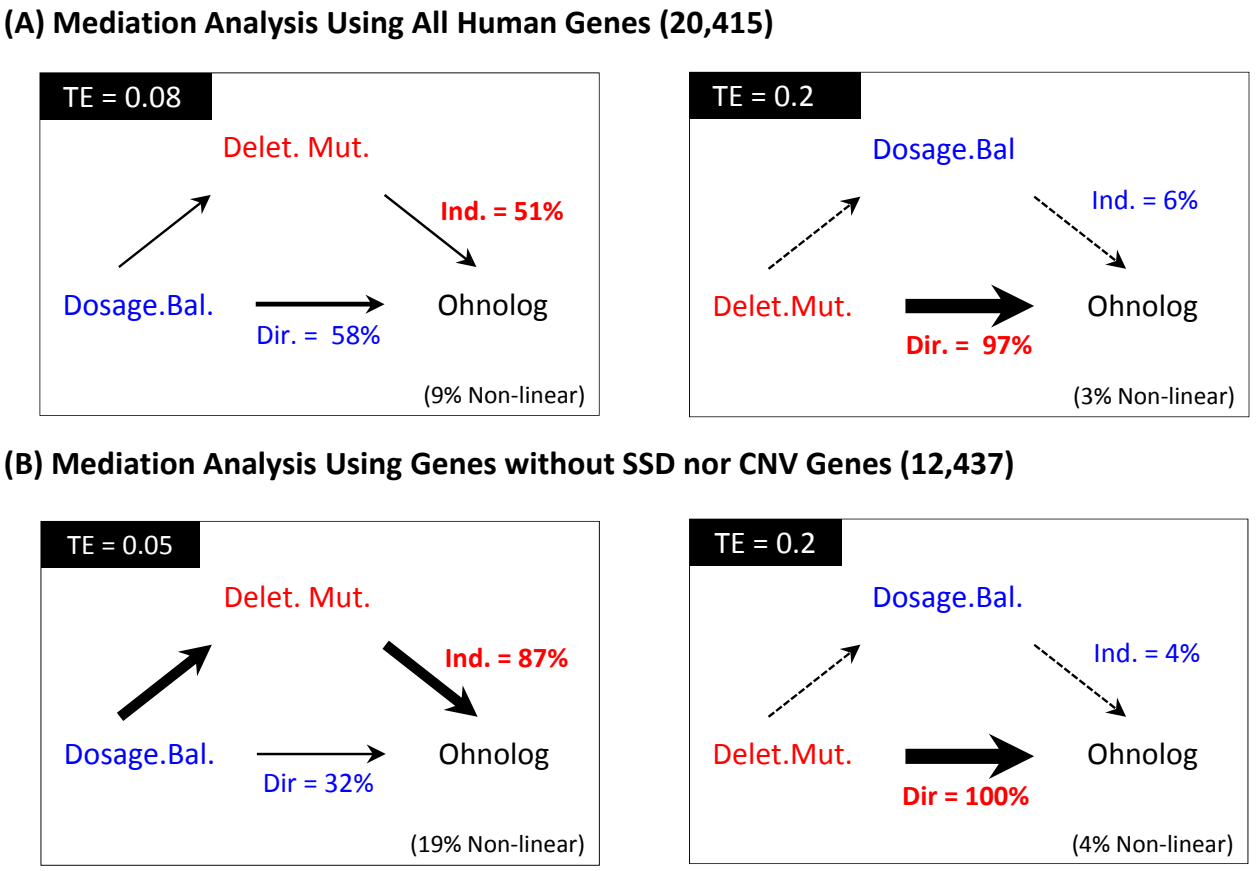

Figure 12.1: Mediation analysis of the indirect effects of deleterious mutations on the retention of ohnologs in multiprotein complexes. Dashed lines represent a small total effect.

complex membership on ohnolog retention with a highly significant mediation by cancer genes ($TE = 0.049, DE/TE = 32.3\%, IE/TE = 91.1\%$, 12.1C), and a strong direct effect of cancer genes without SSD/CNV on ohnolog retention ($TE = 0.227, DE/TE = 99.5\%, IE/TE = 1.7\%$, 12.1D).

We then studied the mediation model involving all protein complexes amd genes susceptible to dominant deleterious mutations (Dom.Mut.). Dom.Mut. included oncogenes, dominant disease genes, dominant negative genes and genes with autoinhibitory folds. We consistently observed a weak total effect of protein complexes on the ohnolog retention and a significant mediation of the total effect by Dom.Mut. genes, whether the SSD/CNV genes are included (12.1E), or not (12.1G). As expected, Dom.Mut. had a very strong direct effect for both gene sets without any mediation of the total effect by protien complexes (12.1F & H).

We observed exactly the same trends with other models restricting ourselves to cancer genes from core dataset and genes susceptible to dominant mutations with oncogenes from core dataset (*core*D.Mut.) as displayed in (Table 12.1 M – P). In fact, the highest total effect with the least mediation on ohnolog retention could be observed for *core*D.Mut. ($TE = 0.262, DE/TE = 100.8\%, IE/TE = 1.9\%$, 12.1P) for all genes excluding SSD/CNV, which is 5 times larger than the total effect of protein complex genes on ohnolog retention for the alternative mediation model (12.1O).

All in all, these observations strongly suggest that dosage balance has a small overall effect on the retention of genes from 2R-WGD. More importantly, a large fraction of this small total effect is in fact indirectly mediated by deleterious mutation susceptibility.

Table 12.1: Summary of mediation analysis for dosage balance & deleterious mutation susceptibility

	Mediation diagram $\nearrow M \searrow$ $X \longrightarrow Y$	**Gene sets:** all genes *versus* w/o SSD & CNV	**Total effect** $X \Longrightarrow Y$ TE	**Direct effect** $X \longrightarrow Y$ DE/TE	**Indirect effect** $X \rightarrow M \rightarrow Y$ IE/TE	Non-linear coupling (% of TE) $DE+IE-TE$
12.1.1	**Delet.Mut.** Dosage.Bal. ⟶ Ohno	all human genes (20,415)	0.086	58.4%	50.7%	9.1%
12.1.2	**Dosage.Bal** Delet.Mut. ⟶ Ohno	all human genes (20,415)	0.200	97.1%	6.1%	3.2%
12.1.3	**Delet.Mut.** Dosage.Bal. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.056	32.3%	87.1%	19.4%
12.1.4	**Dosage.Bal.** Delet.Mut. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.217	100.7%	3.9%	4.6%
A	**Cancer** Complex ⟶ Ohno	all human genes (20,415)	0.078	59.3%	47.7%	7.0%
B	**Complex** Cancer ⟶ Ohno	all human genes (20,415)	0.205	97.9%	4.4%	2.3%
C	**Cancer** Complex ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.049	32.3%	91.1%	23.4%
D	**Complex** Cancer ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.227	101.6%	3.0%	4.6%
E	**Dom.Mut.** Complex ⟶ Ohno	all human genes (20,415)	0.078	73.7%	25.3%	1.0%
F	**Complex** Dom.Mut. ⟶ Ohno	all human genes (20,415)	0.202	96.9%	2.7%	0.4%
G	**Dom.Mut.** Complex. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.049	53.8%	51.6%	5.4%
H	**Complex** Dom.Mut. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.222	99.5%	1.7%	1.1%
I	***core*Cancer** Complex ⟶ Ohno	all human genes (20,415)	0.078	77.9%	28.3%	6.2%
J	**Complex** *core*Cancer ⟶ Ohno	all human genes (20,415)	0.263	99.7%	3.7%	3.4%
K	***core*Cancer** Complex. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.049	64.5%	48.0%	12.5%
L	**Complex** *core*Cancer ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.261	101.8%	2.2%	3.9%
M	***core*D.Mut.** Complex ⟶ Ohno	all human genes (20,415)	0.078	65.0%	39.5%	4.4%
N	**Complex** *core*D.Mut. ⟶ Ohno	all human genes (20,415)	0.278	98.5%	3.6%	2.0%
O	***core*D.Mut.** Complex. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.049	45.4%	63.7%	9.1%
P	**Complex** *core*D.Mut. ⟶ Ohno	all genes w/o SSD & CNV (12,437)	0.262	100.8%	1.9%	2.6%

Dosage.Bal., Dosage Balance; **Delet.Mut.**, Deleterious Mutations; **Ohno**, Ohnologs from relaxed criterion; **Complex**, Protein Complexes; **Cancer**, All cancer genes; **Dom.Mut.**, Dominant Deleterious Mutations; ***core*Cancer**, Cancer genes from core dataset; ***core*D.Mut.**, Dominant Mutations with Core Cancer Genes.

12.1.1 Mediation of 'Dosage.Bal.' ⟹ 'Ohnolog' by 'Delet.Mut.' genes

We analyze below, the association table between the three binary categories: 'Dosage.Bal.', 'Delet.Mut.' and 'Ohnolog' genes to study the total & direct effect of dosage balance constraints on ohnolog retention, and its mediation by deleterious mutation susceptibility.

Dosage.Bal.	**Delet.Mut.**	Ohnolog	n_{xzy}	f_x	g_{xz}	h_x
0	0	0	5694	33.4% Ohnolog	23.8% Ohnolog	**47.1% Delet.Mut.**
0	0	1	1781			
0	**1**	0	3726		44.1% Ohnolog	
0	**1**	1	2935			
1	0	0	1368	42.0% Ohnolog	30.5% Ohnolog	**68.6% Delet.Mut.**
1	0	1	601			
1	**1**	0	2276		47.2% Ohnolog	
1	**1**	1	2034			
6279	10971	7351	20415	36.0% Ohnolog		**53.7% Delet.Mut.**

Hence, we find,

- A significant global effect of Dosage.Bal. on Ohnolog retention: 42.0% *vs* 36.0% ($p = 8.012 \times 10^{-23}$, χ^2 test).
- A strong effect of Dosage.Bal. on Delet.Mut. genes: 68.6% *vs* 53.7% ($p = 5.468 \times 10^{-124}$, χ^2 test).
- A significant effect of Delet.Mut. on Ohnolog retention for Dosage.Bal. genes: 47.2% *vs* 42.0% ($P = 3.553 \times 10^{-12}$, χ^2 test).
- A strong effect of Delet.Mut. on Ohnolog retention for non-Dosage.Bal. genes: 44.1% *vs* 33.4% ($P = 1.358 \times 10^{-76}$, χ^2 test).

The Mediation analysis Pearl [2001, 2012a,b] then yields **a small total effect, *TE*= 0.0860**, with *DE*=0.0502, *IE*=0.0436, and the relative direct and indirect effects below,

↗ **Delet.Mut.** ↘ Dosage.Bal. ⟶ Ohnolog	direct effect Dosage.Bal. ⟶ Ohnolog	indirect effect Dosage.Bal. → **Delet.Mut.** → Ohnolog	non-linear combination of direct and indirect effects
sufficient (as sole cause)	DE/TE =58.3%	IE/TE =50.6%	8.9%
necessary (as complementary cause)	1−IE/TE =49.4%	1−DE/TE =41.7%	

which implies that,

- Only 49.4% of the global effect on Ohnolog retention is owed to the direct link: Dosage.Bal. → Ohnolog (*i.e.* global expected effect if the mediation by Delet.Mut. were 'deactivated').
- 41.7% of the global effect on Ohnolog retention is owed to the indirect mediation by Delet.Mut. (*i.e.* global expected effect if the direct link were 'deactivated').

hence,

- Mediation by Delet.Mut. is sufficient (as sole cause) to account for 50.6% of the Dosage.Bal. ⟹ Ohnolog link.
- Mediation by Delet.Mut. is necessary (as complementary cause) to account for 41.7% of the Dosage.Bal. ⟹ Ohnolog link.

12.1.2 Mediation of 'Delet.Mut.' ⟹ 'Ohnolog' by 'Dosage.Bal.' genes

We analyze below, the association table between the three binary categories, categories: 'Delet.Mut.', 'Dosage.Bal.' and 'Ohnolog' genes to study the total & direct effect of deleterious mutation susceptibility on the ohnolog retention.

Delet.Mut.	**Dosage.Bal.**	Ohnolog	n_{xzy}	f_x	g_{xz}	h_x
0	0	0	5694	25.2% Ohno	23.8% Ohnolog	**20.8% Dosage.Bal.**
0	0	1	1781			
0	**1**	0	1368		30.5% Ohnolog	
0	**1**	1	601			
1	0	0	3726	45.3% Ohnolog	44.1% Ohnolog	**39.3% Dosage.Bal.**
1	0	1	2935			
1	**1**	0	2276		47.2% Ohnolog	
1	**1**	1	2034			
10971	6279	7351	20415	36.0% Ohnolog		**30.8% Dosage.Bal.**

Hence, we find,

- A strong global effect of Delet.Mut. on Ohnolog retention: 45.3% *vs* 36.0% ($p = 2.981 \times 10^{-91}$, χ^2 test).
- A strong effect of Delet.Mut. on Dosage.Bal. genes: 39.3% *vs* 30.8% ($p = 1.774 \times 10^{-83}$, χ^2 test).
- A non-significant effect of Dosage.Bal. on Ohnolog retention for Delet.Mut. genes: 47.2% *vs* 45.3% (0.012, χ^2 test).
- A weak effect of Dosage.Bal. on Ohnolog retention for non-Delet.Mut. genes: 30.5% *vs* 25.2% ($p = 6.093 \times 10^{-9}$, χ^2 test).

The Mediation analysis Pearl [2001, 2012a,b] then yields a **strong total and direct effect, *TE*=0.2007**, *DE*=0.1949, *IE*=0.0123 and the weak indirect effect.

↗ **Dosage.Bal.** ↘ Delet.Mut. ⟶ Ohnolog	direct effect Delet.Mut. ⟶ Ohnolog	indirect effect Delet.Mut. → **Dosage.Bal.** → Ohnolog	non-linear combination of direct and indirect effects
sufficient (as sole cause)	DE/TE =97.1%	IE/TE =6.2%	3.3%
necessary (as complementary cause)	$1-IE/TE$ =93.8%	$1-DE/TE$ =2.9%	

which implies that,

- 93.8% of the global effect on Ohnolog retention is owed to the direct link: Delet.Mut. → Ohnolog (*i.e.* global expected effect if the mediation by Dosage.Bal. were 'deactivated').
- Only 2.9% of the global effect on Ohnolog retention is owed to the indirect mediation by Dosage.Bal. (*i.e.* global expected effect if the direct link were 'deactivated'),

hence,

- Mediation by Dosage.Bal. is sufficient (as sole cause) to account for only 6.2% of the Delet.Mut. ⟹ Ohnolog link.
- Mediation by Dosage.Bal. is necessary (as complementary cause) to account for only 2.9% of the Delet.Mut. ⟹ Ohnolog link.

12.1.3 Mediation of 'Dosage.Bal.' ⟹ 'Ohnolog' by 'Delet.Mut.' genes after excluding SSD and CNV genes

We analyze below, the association table between the three binary categories, categories: 'Dosage.Bal.', 'Delet.Mut.' and 'Ohnolog' genes to study the total & direct effect of dosage balance constraints on ohnolog retention, and its mediation by deleterious mutation susceptibility, restricted to 12,437 protein coding genes after excluding both recent SSD and CNV genes.

Dosage.Bal.	**Delet.Mut.**	Ohnolog	n_{xzy}	f_x	g_{xz}	h_x
0	0	0	2945	39.3% Ohnolog	27.7% Ohnolog	**49.9% Delet.Mut.**
0	0	1	1128			
0	**1**	0	1986		51.0% Ohnolog	
0	**1**	1	2067			
1	0	0	848	45.0% Ohnolog	32.1% Ohnolog	**71.0% Delet.Mut.**
1	0	1	401			
1	**1**	0	1524		50.2% Ohnolog	
1	**1**	1	1538			
4311	7115	5134	12437	41.3% Ohnolog		**57.2% Delet.Mut.**

Hence, we find **after excluding 'SSD+CNV' genes**,

- A weak global effect of Dosage.Bal. on Ohnolog retention: 45.0% *vs* 41.3% ($p = 8.158 \times 10^{-07}$, χ^2 test).
- A strong effect of Dosage.Bal. on Delet.Mut. genes: 71.0% *vs* 57.2% ($p = 4.075 \times 10^{-75}$, χ^2 test).
- A not-so strong effect of Delet.Mut. on Ohnolog retention for Dosage.Bal. genes: 50.2% *vs* 45.0% ($p = 5.205 \times 10^{-9}$, χ^2 test).
- A strong effect of Delet.Mut. on Ohnolog retention for non-Dosage.Bal. genes: 51.0% *vs* 39.3% ($p = 2.432 \times 10^{-52}$, χ^2 test).

The Mediation analysis Pearl [2001, 2012a,b] then yields a **very weak total effect, *TE*=0.0566**, with *DE*=0.0183, *IE*=0.0493 and the relative direct and indirect effects,

↗ **Delet.Mut.** ↘ Dosage.Bal. ⟶ Ohnolog	direct effect Dosage.Bal. ⟶ Ohnolog	indirect effect Dosage.Bal. → **Delet.Mut.** → Ohnolog	non-linear combination of direct and indirect effects
sufficient (as sole cause)	DE/TE =32.3%	IE/TE =87.1%	19.4%
necessary (as complementary cause)	1−IE/TE =12.9%	1−DE/TE =67.7%	

which implies that,

- Only 12.9% of the global effect on Ohnolog retention is owed to the direct link: Dosage.Bal. → Ohnolog (*i.e.* global expected effect if the mediation by Delet.Mut. were 'deactivated').
- 67.7% of the global effect on Ohnolog retention is owed to the indirect mediation by Delet.Mut. (*i.e.* global expected effect if the direct link were 'deactivated').

hence,

- Mediation by Delet.Mut. is sufficient (as sole cause) to account for 87.1% of the Dosage.Bal. ⟹ Ohnolog link.
- Mediation by Delet.Mut. is necessary (as complementary cause) to account for 67.7% of the Dosage.Bal. ⟹ Ohnolog link.

12.1.4 Mediation of 'Delet.Mut.' ⟹ 'Ohnolog' by 'Dosage.Bal.' genes after excluding SSD and CNV genes

We analyze below, the association table between the three binary categories, categories: 'Delet.Mut.', 'Dosage.Bal.' and 'Ohnolog' genes to study the total & direct effect of deleterious mutation susceptibility on the ohnolog retention from 12,437 genes after excluding recent SSDs or CNV genes.

Delet.Mut.	**Dosage.Bal.**	Ohnolog	n_{xzy}	f_x	g_{xz}	h_x
0	0	0	2945	28.7% Ohnolog	27.7% Ohnolog	**23.5% Dosage.Bal.**
0	0	1	1128			
0	**1**	0	848		32.1% Ohnolog	
0	**1**	1	401			
1	0	0	1986	50.7% Ohnolog	51.0% Ohnolog	**43.0% Dosage.Bal.**
1	0	1	2067			
1	**1**	0	1524		50.2% Ohnolog	
1	**1**	1	1538			
7115	4311	5134	12437	41.3% Ohnolog		**34.7% Dosage.Bal.**

Hence, we find **after excluding 'SSD+CNV' genes**,

- A strong global effect of Delet.Mut. on Ohnolog retention: 50.7% *vs* 41.3% ($p = 3.337 \times 10^{-58}$, χ^2 test).
- A strong effect of Delet.Mut. on Dosage.Bal. genes: 43.0% *vs* 34.7% ($p = 7.952 \times 10^{-50}$, χ^2 test).
- An insignificant effect of Dosage.Bal. on Ohnolog retention for Delet.Mut. genes: 50.2% *vs* 50.7% (0.62, χ^2 test).
- A very weak effect of Dosage.Bal. on Ohnolog retention for non-Delet.Mut. genes: 32.1% *vs* 28.7% ($p = 0.008$, χ^2 test).

The Mediation analysis Pearl [2001, 2012a,b] then yields a **very strong total effect, *TE*=0.2194**, with *DE*=0.2209, *IE*=0.0086 and the relative direct and indirect effects,

↗ **Dosage.Bal.** ↘ Delet.Mut. ⟶ Ohnolog	direct effect Delet.Mut. ⟶ Ohnolog	indirect effect Delet.Mut. → **Dosage.Bal.** → Ohnolog	non-linear combination of direct and indirect effects
sufficient (as sole cause)	DE/TE =100.7%	IE/TE =3.9%	4.6%
necessary (as complementary cause)	1−IE/TE =96.1%	1−DE/TE =-0.7%	

which implies that,

- 96.1% of the global effect on Ohnolog retention is owed to the direct link: Delet.Mut. → Ohnolog (*i.e.* global expected effect if the mediation by Dosage.Bal. were 'deactivated').
- -0.7% of the global effect on Ohnolog retention is owed to the indirect mediation by Dosage.Bal. (*i.e.* global expected effect if the direct link were 'deactivated'),

hence,

- Mediation by Dosage.Bal. is sufficient (as sole cause) to account for only 3.9% of the Delet.Mut. ⟹ Ohnolog link.
- Mediation by Dosage.Bal. is necessary (as complementary cause) to account for only -0.7% of the Delet.Mut. ⟹ Ohnolog link.

12.2 Small effect of essentiality on ohnolog retention

Table 12.2: Summary of mediation analysis for essentiality & deleterious mutation susceptibility

	Mediation diagram $\nearrow M \searrow$ $X \longrightarrow Y$	**Gene sets:** all genes *versus* w/o SSD & CNV	**Total effect** $X \Longrightarrow Y$ TE	**Direct effect** $X \longrightarrow Y$ DE/TE	**Indirect effect** $X \rightarrow M \rightarrow Y$ IE/TE	Non-linear coupling (% of TE) $DE+IE-TE$
A	**Delet.Mut.** / *mouse*Ess. ⟶ Ohno	all genes tested for Ess. (6,436)	0.051	70.4%	37.8%	8.2%
B	***mouse*Ess.** / Delet.Mut. ⟶ Ohno	all genes tested for Ess. (6,436)	0.139	96.5%	7.5%	4.0%
C	**Delet.Mut.** / *mouse*Ess. ⟶ Ohno	all genes tested w/o SSD/CNV (4,633)	0.036	61.0%	63.0%	24.0%
D	***mouse*Ess.** / Delet.Mut. ⟶ Ohno	all genes tested w/o SSD/CNV (4,633)	0.145	99.5%	9.0%	8.5%
E	**Delet.Mut.** / *human*Ess. ⟶ Ohno	all genes tested for Ess. (18,938)	0.117	49.6%	45.4%	5.0%
F	***human*Ess.** / Delet.Mut. ⟶ Ohno	all genes tested for Ess. (18,938)	0.179	94.1%	4.3%	1.6%
G	**Delet.Mut.** / *human*Ess. ⟶ Ohno	all genes tested w/o SSD/CNV (11,549)	0.096	46.4%	55.4%	1.8%
H	***human*Ess.** / Delet.Mut. ⟶ Ohno	all genes tested w/o SSD/CNV (11,549)	0.187	96.1%	4.5%	0.5%

Delet.Mut., Deleterious Mutations; **Ohno**, Ohnologs from relaxed criterion; ***mouse*Ess.**, human orthologs of mouse essential genes; ***human*Ess.**, human cell line essential genes; **Ess.**, Essentiality.

The relationship between duplicability and essentiality has remained enigmatic. It has been argued that duplicates confer robustness against null mutations in yeast [Gu et al., 2003] and therefore essential genes are enriched in gene duplicates as compared to genes without any detectable paralog *i.e.* singletons. However, other studies argued that there is no significant association between gene duplicability and essentiality for yeast [He and Zhang, 2006] and mammals [Liang and Li, 2007; Liao and Zhang, 2007]. No distinction between ohnologs and SSD was made in these studies.

Lin et al. [Lin et al., 2007] divided yeast genes into WGD and Non-WGD duplicates and found that, except for ribosomal proteins, most ohnologs in their dataset were dispensable or non-essential. Similar observations have been made by [Hakes et al., 2007; Guan et al., 2007]. In mouse however, Makino et al. observed that essential genes are only enriched in ohnologs but not SSDs [Makino et al., 2009] and attributed this differential retention to dosage balance constraints.

We observed, in agreement with [Makino et al., 2009], that essential genes are slightly enriched in ohnologs. However if genes susceptible to deleterious mutations are removed from essential genes, this enrichment vanishes (Section 10.5.2), leading us to suspect that the effect of essentiality could also be indirectly mediated by mutation susceptibility. To test this, we performed Mediation analysis to study the total, direct and indirect effects of essentiality on ohnolog retention. Essential genes were obtained from two different sources: human orthologs of mouse genes tested for essentiality, and essential genes in human mammary cell lines, as described in Section 7.6.

As depicted in Table 12.2, restricting to the genes tested for essentiality, for both the datasets, we observe that the total effect of essentiality on ohnolog retention is very low. In addition, there is a significant mediation of this total effect by deleterious mutation susceptibility. The statistics becomes stronger if genes duplicated by recent SSD or CNV are removed.

On the other hand, assuming that deleterious mutations play a prominent role in ohnolog retention, we find, in agreement to our expectations that most of the effect of Delet.Mut. is through direct path with three to four time as much total effect.

These trends confirm again that deleterious mutation susceptibility, and not essentiality is the primary cause of ohnolog retention in the human genome.

12.3 Negative causal effect of high expression on ohnolog retention

Table 12.3: Summary of mediation analysis for expression level & deleterious mutation susceptibility

	Mediation diagram $\nearrow M \searrow$ $X \longrightarrow Y$	**Gene sets:** all genes *versus* w/o SSD & CNV	**Total effect** $X \Longrightarrow Y$ TE	**Direct effect** $X \longrightarrow Y$ DE/TE	**Indirect effect** $X \rightarrow M \rightarrow Y$ IE/TE	Non-linear coupling (% of TE) $DE+IE-TE$
A	**Delet.Mut.** Expr.*50p* → Ohno	genes with Expr. value (13,425)	−0.0004	1175.0%	-1075.0%	0%
B	**Expr.*50p*** Delet.Mut. → Ohno	genes with Expr. value (13,425)	0.164	100.1%	-0.1%	0.1%
C	**Delet.Mut.** Expr.*50p* → Ohno	genes with Expr. w/o SSD & CNV (8,651)	−0.0276	113.4%	-14.5%	1.1%
D	**Expr.*50p*** Delet.Mut. → Ohno	genes with Expr. w/o SSD & CNV (8,651)	0.172	100.5%	-0.3%	0.1%
E	**Delet.Mut.** Expr.*75p* → Ohno	genes with Expr. value (13,425)	−0.021	85.1%	14.9%	0%
F	**Expr.*75p*** Delet.Mut. → Ohno	genes with Expr. (20,415)	0.164	99.9%	0.1%	0%
G	**Delet.Mut.** Expr.*75p* → Ohno	genes with Expr. w/o SSD & CNV (8,651)	−0.052	92.1%	9.4%	1.5%
H	**Expr.*75p*** Delet.Mut. → Ohno	genes with Expr. w/o SSD & CNV (8,651)	0.172	99.2%	0.3%	0.5%
I	**Delet.Mut.** Expr.*90p* → Ohno	genes with Expr. value (13,425)	−0.056	88.2%	11.6%	0.2%
J	**Expr.*90p*** Delet.Mut. → Ohno	genes with Expr. value (13,425)	0.164	99.6%	0.5%	0.1%
K	**Delet.Mut.** Expr.*90p* → Ohno	genes with Expr. w/o SSD & CNV (8,651)	0.081	86.0%	14.3%	0.4%
L	**Expr.*90p*** Delet.Mut. → Ohno	genes with Expr. w/o SSD & CNV (8,651)	0.172	98.8%	1.0%	0.1%

Delet.Mut., Deleterious Mutations; **Ohno**, Ohnologs from relaxed criterion; **Expr.**, Expression level; **Expr.*50p***; Median expression; **Expr.*75p***; Expression 75^{th} percentile; **Expr.*90p***; Expression 90^{th} percentile

Another functional genomic property known to be associated with ohnolog retention is the total mRNA levels in cells [Seoighe and Wolfe, 1999; Gout et al., 2009, 2010]. However as described in Section 11.2, we could not observe any significant association between higher expression level and human ohnologs. To confirm the observations from simple statistical association, we also performed Mediation analysis with gene expression level, deleterious mutations and ohnolog retention. We binarize the expression levels at three different percentile points: median, 75^{th} and 90^{th} percentile. Genes having expression levels above these three cutoffs were considered to be highly expressed, or else lowly expressed.

As depicted in Table 12.3, genes having expression levels above the median have a vanishingly small negative effect on ohnolog retention ($TE = -0.0004$, 12.3A). Yet, Delet.Mut. have a very large total effect, all of which is through direct path ($TE = 0.164, DE/TE = 100.1\%, IE/TE = -0.1\%$, 12.3B). The negative effect of expression level on ohnolog retention becomes even

greater for 8,651 genes with expression level without SSD and CNV genes from our dataset (12.3C).

The alternative model assuming that Delet.Mut. is the direct cause of ohnolog retention consistently shows greater total and direct effects after removing SSD/CNV genes (TE = 0.172, DE/TE = 100.5%, IE/TE = −0.3%, 12.3D). The negative total and direct effects of expression level persist and become even stronger for genes having top 25% or top 10% expression levels (12.3 E-G).

These observation confirm that expression level does not play an important role in the retention of human ohnologs. In fact, high expression levels even have very small negative total and direct effect on ohnolog retention, confirming the observed depletion of very highly expressed genes in ohnologs.

12.4 Sequence conservation & ohnolog retention

Ohnologs tend to be highly conserved with low non-synonymous (Ka) to synonymous (Ks) mutation ratios (Ka/Ks) [Brunet et al., 2006; Sémon and Wolfe, 2008]. It has also been suggested that conserved protein are preferentially retained after duplication in the eukaryotic genomes [Davis and Petrov, 2004]. In addition, evolutionary rate has been found to be correlated with ohnolog retention [Jiang et al., 2013]. We also observed that human ohnologs tend to be highly conserved with significantly lower Ka/Ks ratios than non-ohnologs (Section 10.4, [Singh et al., 2012]). We asked whether, using mediation analysis, we can disentangle direct from indirect effects of conservation and deleterious mutations on ohnolog retention, *i.e.* if the low conservation is the prominent cause of the retention of genes after WGD or if it is their mutation susceptibility. In the latter case, the observed high conservation would be just a by-product of the high mutation susceptibility of ohnolog genes.

12.4.1 Mediation with low Ka/Ks values

First we performed mediation among three binary categories: Ka/Ks values from *human-amphioxus* orthologs (15,508 genes with Ka/Ks values), Delet.Mut. and ohnolog retention. Ka/Ks ratios were binarized using three different levels, at median, 25^{th} and 10^{th} percentiles. For each case genes having less Ka/Ks ratios than the three cutoffs were taken to be highly conserved and the rest were considered as not highly conserved.

The results depicted in Table 12.4 show that Ka/Ks ratios do not strongly influence or mediate the effect of deleterious mutations. Using median as a cutoff with 12,607 gene as highly conserved, we observe a positive total and strong direct effect of Ka/Ks on ohnolog retention (TE = 0.097, DE/TE = 89.4%, IE/TE = 8%, 12.4A). This total effect is further reduced after removing SSD/CNV genes from the dataset (TE = 0.073, DE/TE = 87%, IE/TE = 11.2%, 12.4C). The total effect of the alternative Mediation model using Delet.Mut. as primary direct cause is found to be greater than that of Ka/Ks ratios, using all the three cutoffs (TE = 0.139 for all genes, and TE = 0.147 for genes without SSD/CNV). Furthermore, almost all of this total effect is exerted directly on ohnolog retention, without any mediation by sequence conservation.

These trends become even stronger for very highly conserved sequences having Ka/Ks ratios lower than 25^{th} or 10^{th} percentile ranks. Ka/Ks ratios lower than 25^{th} percentile have even lower total effects, 0.054 and 0.030 with and without SSD/CNV respectively (12.4E & G). While the total effect of Delet.Mut. is consistent (12.4F & H). Similarly for 6,460 genes with lowest Ka/Ks, there is a ten times lower total effect of Ka/Ks than that of Delet.Mut. on ohnolog

retention. Furthermore, all the effect of Delet.Mut. on ohnolog retention is direct with no mediation at all (12.4 I–L).

Table 12.4: Summary of mediation analysis for low Ka/Ks ratios from *Amphioxus* & deleterious mutation susceptibility

	Mediation diagram $\nearrow M \searrow$ $X \longrightarrow Y$	**Gene sets:** all genes *versus* w/o SSD & CNV	**Total effect** $X \Longrightarrow Y$ TE	**Direct effect** $X \longrightarrow Y$ DE/TE	**Indirect effect** $X \rightarrow M \rightarrow Y$ IE/TE	Non-linear coupling (% of TE) $DE+IE-TE$
A	**Delet.Mut.** / Ka/Ks*50p* ⟶ Ohno	genes with Ka/Ks value (15,508)	0.097	89.4%	8.0%	2.6%
B	**Ka/K.s*50p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	94.8%	3.4%	1.8%
C	**Delet.Mut.** / Ka/Ks.*50p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.073	87.0%	11.2%	1.8%
D	**Ka/Ks.*50p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	96.7%	2.3%	1.0%
E	**Delet.Mut.** / Ka/Ks.*25p* ⟶ Ohno	genes with Ka/Ks value (15,508)	0.054	76.3%	20.6%	3.2%
F	**Ka/Ks.*25p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	97.7%	1.4%	0.9%
G	**Delet.Mut.** / Ka/Ks.*25p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.030	64.1%	36.5%	0.7%
H	**Ka/Ks.*25p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	99.2%	1.0%	0.1%
I	**Delet.Mut.** / Ka/Ks.*10p* ⟶ Ohno	genes with Ka/Ks value (15,508)	0.010	24.3%	71.8%	3.9%
J	**Ka/Ks.*10p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	99.9%	0%	0.1%
K	**Delet.Mut.** / Ka/Ks.*10p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.004	–76.2%	209.5%	33.3%
L	**Ka/Ks.*10p*** / Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	100.2%	0.2%	0.4%

Delet.Mut., Deleterious Mutations; **Ohno**, Ohnologs from relaxed criterion; **Ka/Ks.*50p***, Ka/Ks lower than median; **Ka/Ks.*25p***, Ka/Ks lower than 25^{th} percentile; **Ka/Ks.*10p***, Ka/Ks lower than 10^{th} percentile

12.4.2 Mediation with high Ka/Ks values

We then analyzed the complementary datasets, to see the effect of high instead of low Ka/Ks values (Table 12.5). As for expression levels, we binarized Ka/Ks at median, 75^{th} and 90^{th} percentile ranks, and investigated whether the genes having high Ka/Ks could be the primary direct cause of ohnolog retention. As expected, for the genes having Ka/Ks higher than median (Complement of Table 12.4 A–D), we observed that there is an equal but negative total effect of high Ka/Ks on ohnolog retention, while the total effect of Delet.Mut. remains unchanged (Table 12.5 A–D).

For higher Ka/Ks values while the total and direct effect of Delet.Mut. on ohnolog retention is very strong (Table 12.5 F, H J & L), we observe interestingly, a very high negative total effect of Ka/Ks. For example, genes with Ka/Ks greater than 75^{th} percentile have a very strong negative effect on ohnolog retention, $TE = -0.181$ with all genes and $TE = -0.132$ without SSD/CNV. Majority of this effect is direct. Furthermore, genes which have tolerated a very high non-synonymous mutations, with highest Ka/Ks values (*i.e.* positive selection) an even stronger total negative effect, –0.204 and –0.152 with all genes or excluding SSD/CNV, respectively (12.5 I–K).

Table 12.5: Summary of mediation analysis for high Ka/Ks ratios from *Amphioxus* & deleterious mutation susceptibility

	Mediation diagram $\nearrow M \searrow$ $X \longrightarrow Y$	**Gene sets:** all genes *versus* w/o SSD & CNV	**Total effect** $X \Longrightarrow Y$ TE	**Direct effect** $X \longrightarrow Y$ DE/TE	**Indirect effect** $X \rightarrow M \rightarrow Y$ IE/TE	Non-linear coupling (% of TE) $DE+IE-TE$
A	**Delet.Mut.** Ka/Ks*50p* ⟶ Ohno	genes with Ka/Ks value (15,508)	– 0.097	92.0%	10.5%	2.5%
B	**Ka/K.s*50p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	94.8%	3.4%	1.8%
C	**Delet.Mut.** Ka/Ks.*50p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	–0.073	88.8%	13.2%	1.9%
D	**Ka/Ks.*50p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	96.7%	2.3%	1.0%
E	**Delet.Mut.** Ka/Ks.*75p* ⟶ Ohno	genes with Ka/Ks value (15,508)	– 0.181	97.0%	6.3%	3.2%
F	**Ka/Ks.*75p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	91.2%	5.5%	3.2%
G	**Delet.Mut.** Ka/Ks.*75p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	– 0.132	94.8%	8.9%	3.6%
H	**Ka/Ks.*75p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	94.5%	3.1%	2.4%
I	**Delet.Mut.** Ka/Ks.*90p* ⟶ Ohno	genes with Ka/Ks value (15,508)	– 0.204	97.7%	4.7%	2.4%
J	**Ka/Ks.*90p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks value (15,508)	0.139	95.8%	2.9%	1.3%
K	**Delet.Mut.** Ka/Ks.*90p* ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	– 0.152	92.7%	9.3%	2.0%
L	**Ka/Ks.*90p*** Delet.Mut. ⟶ Ohno	genes with Ka/Ks w/o SSD & CNV (9,850)	0.147	96.8%	2.5%	0.7%

Delet.Mut., Deleterious Mutations; **Ohno**, Ohnologs from relaxed criterion; **Ka/Ks.*50p***, Ka/Ks lower than median; **Ka/Ks.*25p***, Ka/Ks lower than 25^{th} percentile; **Ka/Ks.*10p***, Ka/Ks lower than 10^{th} percentile

This suggests that although conserved sequences have a weaker relation with ohnolog retention, rapidly evolving sequences have retained less ohnologs. Sequence that have tolerated many mutations even oppose the retention of ohnologs. Although higher Ka/Ks ratio is not a perfect measure of positive selection, these observation hints towards a lack of positive selection in ohnologs, likely owing to population bottleneck in post WGD populations (See Discussion).

All in all, the results from mediation analysis strongly suggest that — (1) Deleterious mutation susceptibility has the strongest causal effect on ohnolog retention. (2) All the other studied genomic attributes have relatively weaker causal effects on ohnolog retention. (3) Almost all of the effect of deleterious mutation susceptibility is through a direct path, with negligible mediation from other properties. (4) The effect of other genomic attributes is mostly mediated by the deleterious mutation susceptibility.

13

Population Genetic Model for the Retention of "Dangerous" Ohnologs

As demonstrated in previous chapters, human genes with a documented sensitivity to dominant deleterious mutations have retained statistically more ohnologs from the two WGD events at the onset of jawed vertebrates. This suggests that ohnologs have been retained in the vertebrate genomes, not because they initially brought selective advantages following WGD, but because their mutations were more likely detrimental or lethal than nonfunctional, thereby preventing their rapid elimination from the genomes of surviving individuals following WGD transitions, as outlined in the evolutionary model depicted in Figure 13.1. For completeness and clarity, Figure 13.1 examines all possible evolutionary scenarios following either a SSD or a WGD duplication event in the genome of one or a few individuals in an initial population.

The first and critical difference between SSD and WGD duplication events occurs at the population genetics level with an obligate speciation following WGD event, owing to the difference in ploidy between pre- and post-WGD individuals. As a result, all individuals in the post-WGD population carry twice as many genes as their pre-WGD relatives, whereas only a few individuals in the post-SSD population carry a single small duplicated region. Figure 13.1 then outlines the three mutation/selection scenarios focusing on a single gene duplicate in the genomes of post-SSD or post-WGD populations:

(A) Beneficial mutations after SSD or WGD are expected to spread and become eventually fixed in the new populations, although the bottleneck in population size following WGD limits in practice the efficacy of adaptation in post-WGD species.

(B) Neutral or nearly neutral mutations mainly lead to the random nonfunctionalization of one copy of most redundant gene duplicates and, therefore, to their elimination following both SSD and WGD events. In post- WGD populations, this results in the "reciprocal gene loss" of most gene duplicates, which is also known to lead to further speciations in post-WGD species, owing to the interbreeding incompatibility between post-WGD individuals with different "reciprocal gene loss" pattern [Lynch and Force, 2000b]. Alternatively, neutral or nearly neutral mutations can also result in the eventual retention of both duplicate

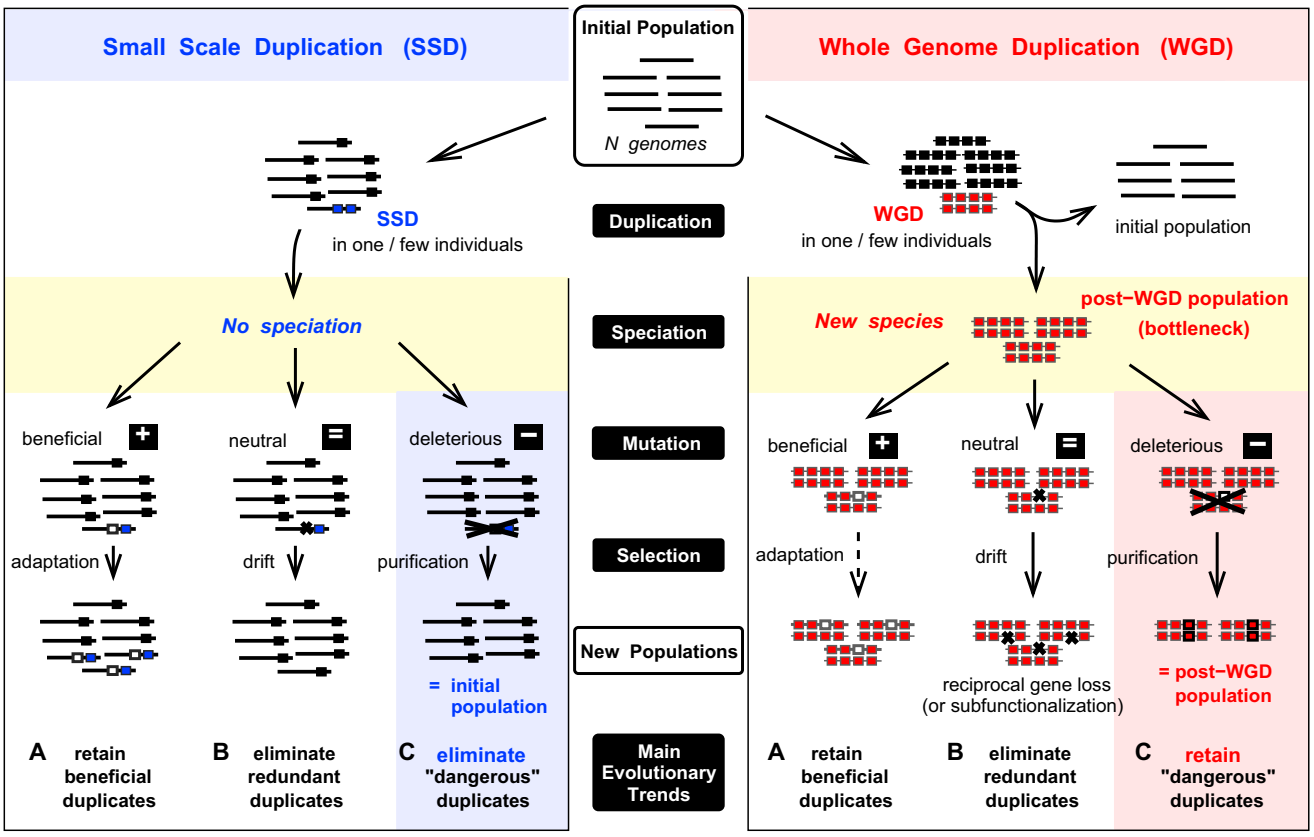

Figure 13.1: Evolutionary trends of duplicated genes following SSD or WGD.
(A–C) Horizontal lines represent the genome of different individuals. Square blocks symbolize the genes, duplicated (SSD: blue; WGD: red) or not (black). Black crosses highlight the loss of one gene (small crosses) or the elimination of an individual (larger crosses), whereas bordered square blocks emphasize retained mutated copies. Evolutionary scenarios are depicted at the population genetics level following either a SSD (left panel) or a WGD (right panel) in one or a few individuals of an initial population. Unlike SSD, WGD is invariably coupled to a speciation event, owing to the difference in ploidy between pre- and post-WGD individuals. Three possible scenarios—beneficial (A), neutral or nearly neutral (B), or deleterious mutations (C) in one gene duplicate—are outlined in post-SSD and post-WGD populations. The main difference concerns the mutation/selection process of the genes prone to dominant deleterious mutations (C).

copies through subfunctionalization [Hughes, 1994; Lynch and Force, 2000a], that is, by rendering each duplicate copy unable to perform all the functions of their ancestral gene.

(C) Finally, dominant deleterious mutations favor the elimination of the individuals (or their descendants) harboring them through purifying selection. However, this typically leads to opposite outcomes in post-SSD and post-WGD populations. In post-SSD populations, dominant deleterious mutations will tend to eliminate SSD duplicates before they have the time to reach fixation (see below). By contrast, in post-WGD populations, where all ohnologs have been initially fixed through WGD-induced speciation, purifying selection will effectively favor the retention of dangerous ohnologs, as all surviving individuals still present (non-deleterious) functional copies of these dangerous genes.

Note, in particular, that this somewhat counterintuitive evolutionary model for the retention of “dangerous” ohnologs hinges on two unique features:

1. It requires an autosomal dominance of deleterious mutations, in agreement with our observation, above, that retained ohnologs are more “dangerous” than “essential.”

2. It relies on the fact that successful WGD events start with a concomitant speciation event, which immediately fixes all ohnolog duplicates in the initial post-WGD population Figure 13.1.

Note, also, that the same evolutionary trend is expected for dangerous SSD duplicates that would have the time (t) to become fixed through genetic drift in a population of size N before deleterious mutations can arise at a rate K, *i.e.*, $t = 4N < 1/K$. This corresponds to a population bottleneck effect with $N < 1/(4K) \simeq 5,000–10,000$ for typical vertebrates.

Part IV

Discussion & Perspectives

14
Discussion & Perspectives

BEYOND the human and vertebrate genomes, WGD events have now been established in all major eukaryote kingdoms [Sémon and Wolfe, 2007; Evlampiev and Isambert, 2007]. Unlike SSD events, WGD transitions provide a unique evolutionary mechanism, enabling the simultaneous duplication of entire genetic pathways and multiprotein complexes, followed by long periods of functional divergence and extensive loss of ohnologs [Aury et al., 2006]. Moreover, although both WGD and SSD events have expanded the gene repertoires and resulting protein networks [Evlampiev and Isambert, 2007, 2008] of eukaryotes, it has become increasingly clear that WGD and SSD events actually lead to the expansion of different gene classes in the course of evolution, [Maere et al., 2005; Aury et al., 2006; Sémon and Wolfe, 2007; Freeling, 2009; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; Singh et al., 2012].

In this thesis, we report that WGD have effectively favored the expansion of gene families prone to deleterious mutations in the human genome, such as Mendelian disease genes, genes implicated in cancer and genes with autoinhibitory interactions. Furthermore, we also noticed that SSD do not present any significant bias towards the retention of these genes susceptible to dominant deleterious mutations. Our observations clearly highlight the need to discriminate between different duplication mechanisms while analyzing the evolution by gene duplication, to avoid erroneous conclusions. For example, in recent study, [Chen et al., 2013b] have analyzed the evolution of human monogenic disease genes (MD) and concluded that human MD genes have more duplicates (SSDs and ohnologs lumped together) than singletons. We noticed in our analysis, however, that only ohnologs and not SSDs are enriched in disease genes. Furthermore, this bias only concerns dominant disease genes, as recessive disease genes do not exhibit any biased retention of ohnologs, SSD or singletons.

Using causal inference analysis and the Mediation framework, we also showed that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations. From a broader perspective, a number of studies have now shown that many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent. In the light of the present study, we expect that many of these statistically significant correlations mainly result from indirect rather than direct associations, as observed for dosage balance, expression level, essentiality and Ka/Ks ratio in our

analysis. This highlights the need to rely on more advanced inference methods to analyze the multiple, *direct*, and *indirect* causes underlying the evolution of specific gene repertoires. Our results are also clearly supported by partial correlation analysis [Singh et al., 2012], however, the two approaches are not equivalent. In particular, while mediation effects require partial correlation, partial correlation does not imply mediation in general and can exist in absence of mediation [Singh et al., 2012].

To explain our observations we have proposed a simple evolutionary mechanism from a population genetic perspective. We argue that the biased retention of genes susceptible to dominant disease genes by WGD is a consequence of WGD-induced speciation and subsequent purifying selection in the post WGD population. With this sympatric speciation event, all the genes are already fixed in the population to start with and during subsequent evolution most genes prone to loss of function mutations are lost. On the contrary, there is typically no such speciation event coupled to SSD events. Therefore, an SSD duplicate has to rise in frequency in the population to reach fixation. Mutations provide the engine for evolution, and since, most of the mutations in oncogenes, dominant disease genes and genes with autoinhibitory folds are expected to be deleterious, the loss of these genes after WGD and their fixation after SSD would be difficult.

Our schematic model presented in Figure 13.1, is not only strongly supported by the genomic data, but also by a proper population genetics model [Malaguti et al., 2013]. In this analysis, first a deterministic haploid model including initial duplicated loci has been solved, to study the retention of duplicated genes through sub-functionalization against their neutral loss-of-function or deleterious gain-of-function at one locus. Then, to go beyond deterministic solutions for large populations, the formalism of one-step-process master equations and stochastic simulations have been used to analyze the effect of finite population sizes on the retention of SSD *versus* WGD duplicates. The deterministic solution predicts that the fixation of individual genes through WGD *versus* SSD duplicates depends on the rate of dominant deleterious mutations against the total mutation rate. Stochastic simulations are in strong agreement with the predictions of the deterministic solution, suggesting that while deleterious gain-of-function mutations favor the elimination of "dangerous" gene duplicates after SSD, a significantly enhanced retention of "dangerous" duplicates is predicted after WGD. All in all, this population genetics model supports the idea that the enhanced retention of "dangerous" WGD duplicates prone to dominant deleterious mutations is an indirect consequence of the initial speciation events triggered by WGD and the ensuing purifying selection in post-WGD species.

On longer timescales, we expect that this initial retention bias of "dangerous" ohnologs through WGD effectively promote a prolonged genetic drift and, thus, a progressive functional divergence between ohnolog pairs. This eventually favors the subfunctionalization of ancestral functions between ohnolog pairs, which ultimately warrants their long-term maintenance following WGD events. This subfunctionalization process requires, however, that the expression of ohnologs is not rapidly suppressed by large-scale deletion or silencing mutations in regulatory regions. As ohnolog pairs are not arranged in tandem, large-scale deletions through unequal crossing-over cannot typically remove entire ohnolog duplicates while preserving the integrity of nearby genes. Furthermore, as the size of promoter or enhancer regions is typically much smaller than UTRs and coding regions, one expects that the rate of transcriptional silencing does not exceed the rates of functional silencing and divergence in UTRs and coding regions. In fact, early estimates [Nadeau and Sankoff, 1997] showed that gene loss and functional divergence after genome duplications in early vertebrates occurred at comparable rates in gene families including at least two ohnologs. This is also directly evidenced by pseudote-

traploid species like the vertebrate *Xenopus laevis*, which still retains about ~ 40% of its initial ohnologs from a 30-million-year-old WGD [Sémon and Wolfe, 2008]. Theses observations suggests that ohnologs prone to dominant deleterious mutations have at least a few million years to diverge and become nonredundant genes before they have a chance to be deleted or transcriptionally silenced.

Yet, we found that the retention of these dangerous ohnologs remains intrinsically stochastic by nature as many of them have also been eliminated following WGD events. This presumably occurred through loss-of-function mutations, transcriptional silencing, or large-scale deletion before ohnolog pairs could diverge and become nonredundant genes. More quantitatively, a simple theoretical estimate, derived from the long-term retention statistics [Singh et al., 2012] shows that only 6%–10% of the initial ohnolog duplicates have been retained on average at each round of WGD. By comparison, ~ 23%–30% of the initial ohnologs prone to gain-of-function mutations have been retained on average at each WGD. This implies that genes susceptible to deleterious mutations are two to five times more likely to retain ohnologs on long evolutionary timescales. Moreover, genes combining several factors associated with enhanced susceptibility to autosomal-dominant deleterious mutations have been found to be more than ten times more likely to retain ohnologs than genes lacking gain-of-function mutations, as illustrated on the examples of oncogenes with autoinhibitory folds (Figures 10.1 and 10.2).

A small number of genes susceptible to dominant deleterious mutations have also been fixed by SSDs in the human genome. This might come about by positive selection as predicted by our model [Malaguti et al., 2013], however, the possibility of beneficial gain-of-function mutations is very low owing to their susceptibility to deleterious mutations. This is evidenced by their very low retention after SSD in the human genome. In general, the fixation of SSDs by subfunctionalization is also less likely in the absence of positive selection.

There are specific evidences for models of subfunctionalization such as Escape from Adaptive Conflict [Des Marais and Rausher, 2008] and Innovation-Amplification-Divergence [Näsvall et al., 2012] for limited cases. Yet, their role on a genome-wide scale is hard to explain. In general, as the subfunctionalization model was developed before the SSD *versus* WGD bias was reported, it does not explain the strong bias observed by us and many other studies. Indeed, the suspicion that subfunctionalization alone is the major mechanism of duplicated gene retention is growing [MacCarthy and Bergman, 2007; Freeling, 2008, 2009]. The 'true' instances of neofunctionalization are rare. Nevertheless, neofunctionalization also fails to provide an explanation of the biased retention of dominant disease genes by WGD.

The elimination of ohnologs has been shown to drive further speciation events within post-WGD (sub)populations, due to the emergence of recombination barriers from the accumulation of differences in ohnolog deletion patterns between post-WGD individuals [Lynch and Force, 2000b]. The resulting fragmentation of post-WGD subpopulations is then expected to sustain negative selection pressure that favors the retention of the remaining ohnolog pairs prone to deleterious mutations, as outlined in Figure 13.1. Hence, although most WGDs are unlikely to bring much fitness benefit on short evolutionary timescales (if only due to the population bottlenecks associated with WGD-induced speciations), they provide a unique evolutionary mechanism to experiment virtually unlimited combinations of regulation/deletion patterns from redundant ohnolog genes. Over long timescales (>100–500 MY), such trial and error combinations have visibly led to the evolutionary success and radiation of WGD species.

Some previous studies have alluded to the origin of Mendelian disease genes [Makino and McLysaght, 2010; Dickerson and Robertson, 2011] by WGD. In addition to Mendelian disease genes, we also wanted to characterize the genes implicated in cancers, and therefore, cancer

genes and oncogenes contribute a significant fraction to our disease gene dataset. However, the role of cancer mutations in evolution is considered to be insignificant because of the nature of these mutations, which are mostly somatic and are believed to typically occur later in life past the reproductive periods of the extant animals. Furthermore, as the occurrence of cancer has certainly increased in the modern times due to a variety of factors, there is a debate on antiquity of cancer as a disease [Capasso, 2005; David and Zimmerman, 2010a,b; Faltas, 2010; Wang et al., 2010b]. In our analysis however, we observed that dominant cancer genes also present very strong retention bias after WGD. Although, most cancers occur later in the lifespan, cancer is one of the leading cause of death, even in young adults (see Table 6 in [Siegel et al., 2013]). For example, cancer is the first leading cause of natural deaths in both males and females of age 1-19 and 20-39 according to the cancer mortality statistics in Unites States of America in 2009 [Siegel et al., 2013]. Although, such systematic analysis is difficult to obtain for many countries in the world and for other organisms and specially for the ancestors the vertebrates. Yet, even a relatively small rate of cancer instances in juvenile and young organisms in the ancestor of vertebrates would be enough to strongly bias the evolution of genes implicated in cancers on a time-scale of 100-500 MY. In addition, mutations in the germline can also predispose subsequent generations to cancer.

A considerable part of our analysis concerns the accurate detection of ohnologs in the vertebrate genomes. As shown by our results, the *P-value* based approach is indeed efficient in removing possible false positive candidates. Most importantly, since our *P-values* depend on the genomic context, orthologs and paralogs, it allows to bypass arbitrary criteria while detecting ohnologs. However, there are many scopes of improvements in our approach.

We resort to a content based synteny as the conservation of gene collinearity is limited owing to a very long time period of evolution since the 2R-WGD [Putnam et al., 2008]. Yet, our approach could be improved to seek for such regions of conserved gene order and provide more weight to the ohnologs residing on such regions. After WGD, since the ohnologs are lost randomly from the duplicated sister region, each region in an outgroup genome should *alternatively* correspond to multiple duplicated sister regions in the vertebrate genome. Such *alternative* synteny correspondence has been used to prove WGD events in *Saccharomyces cerevisiae* [Kellis et al., 2004] and *Tetraodon* [Jaillon et al., 2004]. Such alternation could be difficult to find in the vertebrate genomes. Yet, if observed in some genomic regions, it would be a tell-tail sign of 'true' ohnologs belonging to such regions.

There is unmistakable evidence now that vertebrates have descended from polyploid ancestors [Dehal and Boore, 2005; Putnam et al., 2008] and our aim in this study was to detect the ohnologs in the vertebrate genomes with high confidence. Therefore, we did not attempt to find an optimum number of ohnologs or the patterns reminiscent of WGD to prove that WGD actually occurred. Yet, we observed that in the genome of organisms such as *Drosophila* and *C. elegans* our approach could not identify many ohnologs. Therefore, with the identification and optimization of minimum required ohnologs and ohnolog families and their genomic distribution, our approach can also be used to detect unknown WGD events in the sequenced (or upcoming) genomes. These improvements would be specially useful to study relatively recent genome duplications events.

We noticed that some of the limitations in detecting the ohnologs are sorting the candidates with correct duplication time and the level of genome annotation. In future, we would like to identify ohnologs in all the sequenced vertebrate and Teleost Fish genome in Ensembl to systematically study the fraction of ohnologs not detected by our approach correspond to the lineage specific loss of ohnologs in different vertebrates or the annotation anomalies. More importantly, our work also opens avenues for a large number of possible analyses such as

the relative impact of 2R- and 3R-WGDs in the fish genomes, attempting to distinguish the genes retained from the 1^{st} or the 2^{nd} round of duplication based on syntenic patterns, and identifying ohnologs in other classes of genes such as non-coding RNAs.

Another important aspect of our work is related to the use of advanced inference methods such as the Mediation analysis which allowed us to study three-property causal relationships. However, a desirable improvement would be to detect causal networks using genomic and functional data on a larger number of variables.

In conclusion, we present evidence supporting an evolutionary link between the susceptibility of human genes to dominant deleterious mutations and the expansion of these "dangerous" gene families by two WGD events at the onset of vertebrates. We propose that deleterious mutations, responsible for many cancers and other severe genetic diseases have also underlain purifying selection over long evolutionary time-scales, which effectively favored the retention of vertebrate ohnologs prone to dominant deleterious mutations. From a population genetics perspective, we argue that this counterintuitive retention of dangerous ohnologs hinges in fact on WGD-induced speciation events, which are largely credited for the genetic complexity and successful radiation of vertebrate species. These findings highlight the importance of purifying selection from WGD events on the evolution of vertebrates and, beyond, exemplify the role of non-adaptive forces on the emergence of eukaryote complexity.

A
Articles

Cell Reports
Article

On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates

Param Priya Singh,[1,3] Séverine Affeldt,[1,3] Ilaria Cascone,[2] Rasim Selimoglu,[2] Jacques Camonis,[2] and Hervé Isambert[1,*]
[1]CNRS UMR168
[2]INSERM U830
UPMC, Institut Curie, Research Center, 26, rue d'Ulm, 75248 Paris, France
[3]These authors contributed equally to this work
*Correspondence: herve.isambert@curie.fr
http://dx.doi.org/10.1016/j.celrep.2012.09.034

SUMMARY

The emergence and evolutionary expansion of gene families implicated in cancers and other severe genetic diseases is an evolutionary oddity from a natural selection perspective. Here, we show that gene families prone to deleterious mutations in the human genome have been preferentially expanded by the retention of "ohnolog" genes from two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates. We further demonstrate that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations. This enhanced retention of "dangerous" ohnologs, defined as prone to autosomal-dominant deleterious mutations, is shown to be a consequence of WGD-induced speciation and the ensuing purifying selection in post-WGD species. These findings highlight the importance of WGD-induced nonadaptive selection for the emergence of vertebrate complexity, while rationalizing, from an evolutionary perspective, the expansion of gene families frequently implicated in genetic disorders and cancers.

INTRODUCTION

Just as some genes happen to be more "essential," owing to their deleterious loss-of-function or null mutations, some genes can be classified as more "dangerous," due to their propensity to acquire deleterious gain-of-function mutations. This is, in particular, the case for oncogenes and genes with autoinhibitory protein folds, whose mutations typically lead to constitutively active mutants with dominant deleterious phenotypes (Pufall and Graves, 2002).

Dominant deleterious mutations, that are lethal or drastically reduce fitness over the lifespan of organisms, must have also impacted their long term evolution on timescales relevant for genome evolution (e.g., >10–100 million years [MY]). In fact, dominant disease genes in humans have been shown to be under strong purifying selection (Furney et al., 2006; Blekhman et al., 2008; Cai et al., 2009). Yet, "dangerous" gene families implicated in cancer and severe genetic diseases have also been greatly expanded by duplication in the course of vertebrate evolution. For example, the single orthologous locus, *Ras85D* in flies and *Let-60* in nematodes, has been duplicated into three *RAS* loci in typical vertebrates, *KRAS*, *HRAS*, and *NRAS*, that present permanently activating mutations in 20%–25% of all human tumors, even though *HRAS* and *NRAS* have also been shown to be dispensable for mouse growth and development (Ise et al., 2000; Esteban et al., 2001).

While the maintenance of essential genes is ensured by their lethal null mutations, the expansion of dangerous gene families remains an evolutionary puzzle from a natural selection perspective. Indeed, considering that many vertebrate disease genes are phylogenetically ancient (Domazet-Loso and Tautz, 2008; Cai et al., 2009; Dickerson and Robertson, 2012), and that their orthologs also cause severe genetic disorders in extant invertebrates (Berry et al., 1997; Ciocan et al., 2006; Robert, 2010), it is surprising that dangerous gene families have been duplicated more than other vertebrate genes without known dominant deleterious mutations. While gene duplicates can confer mutational robustness against loss-of-function mutations, multiple copies of genes prone to gain-of-function mutations are expected to lead to an overall aggravation of a species' susceptibility to genetic diseases and thus be opposed by purifying selection.

Two alternative hypotheses can be put forward to account for the surprising expansion of dangerous gene families. Either, the propensity of certain genes to acquire dominant deleterious mutations could be a mere by-product of their presumed advantageous functions. In that case, only the overall benefit of gene family expansion should matter, irrespective of the mechanism of gene duplication. Alternatively, gene susceptibility to dominant deleterious mutations could have played a driving role in the striking expansion of dangerous gene families. But what could have been the selection mechanism?

In this article, we report converging evidences supporting the latter hypothesis and propose a simple evolutionary model to explain the expansion of such dangerous gene families. It is based on the observation that the majority of human genes prone to dominant deleterious mutations, such as oncogenes and genes with autoinhibitory protein folds, have not been duplicated through small scale duplication (SSD). Instead, the expansion of these dangerous gene families can be traced back to two rounds of whole-genome duplication (WGD), that occurred at the

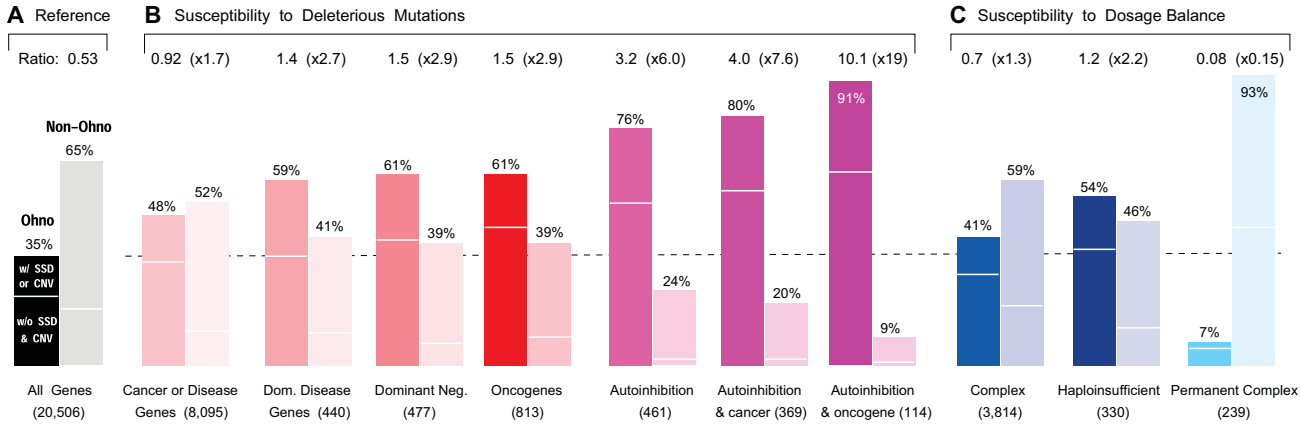

Figure 1. Prevalence of Retained Ohnologs in the Human Genome within Different Gene Classes

(A and B) Prevalence of retained ohnologs either "w/ SSD or CNV" or "w/o SSD & CNV" for all 20,506 human protein-coding genes (A), and gene classes susceptible to deleterious mutations (B). Note that gene classes with higher susceptibility to deleterious mutations retained more ohnologs.

(C) Ohnolog retention in gene classes susceptible to dosage balance constraints. Fold changes in ohnolog/nonohnolog ratios are given relative to the reference from all human genes in (A).

See also Figure S1.

onset of jawed vertebrates, some 500 MY ago (Ohno, 1970; Putnam et al., 2008).

These two rounds of WGD in the early vertebrate lineage are frequently credited with creating the conditions for the evolution of vertebrate complexity. Indeed, WGD-duplicated genes, so-called "ohnologs" in honor of Susumu Ohno (Ohno, 1970; Wolfe, 2000), have been preferentially retained in specific gene classes associated with organismal complexity, such as signal transduction pathways, transcription networks, and developmental genes (Maere et al., 2005; Blomme et al., 2006; Freeling and Thomas, 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010). By contrast, gene duplicates coming from SSD are strongly biased toward different functional categories, such as antigen processing, immune response, and metabolism (Huminiecki and Heldin, 2010). SSD paralogs and WGD ohnologs also differ in their gene expression and protein network properties (Hakes et al., 2007; Guan et al., 2007). Furthermore, recent genome-wide analysis have shown that ohnologs in the human genome have experienced fewer SSD than "nonohnolog" genes and tend to be refractory to copy number variation (CNV) caused by polymorphism of small segmental duplications in human populations (Makino and McLysaght, 2010). These antagonist retention patterns of WGD and SSD/CNV gene duplicates in the human genome have been suggested to result from dosage balance constraints (Makino and McLysaght, 2010) on the relative expressions of multiple protein partners (Veitia, 2002), as proposed earlier for other organisms like yeast (Papp et al., 2003) and the paramecium (Aury et al., 2006).

In this article, we investigate the evolutionary causes responsible for the expansion of gene families prone to deleterious mutations in vertebrates and propose a simple evolutionary model accounting for their antagonistic retention pattern after WGD and SSD events. The retention of ohnologs in the human genome is shown to be more strongly associated with their susceptibility to deleterious mutations, than their functional importance or "essentiality." We also demonstrate using a causal inference analysis, that the retention of many ohnologs suspected to be dosage balanced is in fact an indirect effect of their higher susceptibility to deleterious mutations. We argue that the enhanced retention of dangerous ohnologs is a somewhat counterintuitive yet simple consequence of the speciation event triggered by WGD and the ensuing purifying selection in post-WGD species.

These findings rationalize, from an evolutionary perspective, the WGD expansion of gene families frequently implicated in genetic disorders, such as cancer, and highlight the importance of nonadaptive selection on the emergence of vertebrate complexity.

RESULTS

Genes Prone to Deleterious Mutations Retain More Ohnologs

We first analyzed a possible association between the susceptibility of human genes to deleterious mutations and their retention of ohnologs, as proposed in Gibson and Spring (1998) for multidomain proteins. To this end, we considered multiple classes of genes susceptible to deleterious mutations from experimentally verified databases and literature. These classes include cancer genes (from multiple sources including COSMIC [Forbes et al., 2011] and CancerGenes [Higgins et al., 2007]), genes mutated in other genetic disorders, dominant negative genes from OMIM, and genes with autoinhibitory protein folds (Experimental Procedures). We looked at the relative contributions of WGD and SSD in the expansion of these "dangerous" gene classes.

The results, depicted in Figures 1 and S1, demonstrate indeed a strong association between the retention of human ohnologs from vertebrate WGD and their reported susceptibility to deleterious mutations, as compared to nonohnologs, whereas an

opposite pattern is found for SSD/CNV gene duplicates. Overall, the 8,095 human genes associated with the occurrence of cancer and other genetic diseases have retained significantly more ohnologs than expected by chance, 48% versus 35% (48%; 3,844/8,095; $p = 1.3 \times 10^{-129}$, χ^2 test). Furthermore, these associations, which do not take into account the actual severity of the gene mutations, are clearly enhanced when the analysis is restricted to genes with direct experimental evidence of dominant deleterious mutations, such as dominant disease genes (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), dominant negative mutants (61%; 292/477; $p = 3.9 \times 10^{-34}$, χ^2 test), oncogenes (61%; 493/813; $p = 1.4 \times 10^{-54}$, χ^2 test), or genes exhibiting autoinhibitory constraints (76%; 350/461; $p = 2.7 \times 10^{-77}$, χ^2 test). The biased retention of ohnologs is even stronger for genes combining several factors associated with an enhanced susceptibility to deleterious mutations, such as cancer genes with autoinhibitory folds, (80%; 294/369; $p = 1.0 \times 10^{-73}$, χ^2 test), or oncogenes with autoinhibitory folds, (91%; 104/114; $p = 6.9 \times 10^{-37}$, χ^2 test).

This retention of dangerous ohnologs is illustrated on Table 1 that presents an up-to-date list of 76 hand-curated gene families of up to four ohnologs, exhibiting both autoinhibitory folds and oncogenic properties (see Table S1 for oncogenic and autoinhibitory details and references). These dangerous ohnologs are typically found along signal transduction cascades, from receptor tyrosine kinases and cytoplasmic or nuclear kinases to guanine exchange factors (GEF), GTPase activating proteins (GAP), and transcription factors (Table 1, gene classes A–E). In addition, autoinhibited oncogenes are also found in other ohnolog families with diverse functions (Table 1, gene class F). By contrast, we obtained a hand-curated list of only ten nonohnolog genes exhibiting both autoinhibitory and oncogenic properties, Table 1, gene class G (see Table S2 for oncogenic and autoinhibitory details and references). Interestingly, half of them (4/10) can be traced back to SSD events, which occurred after or at the same period of the two WGD in early vertebrate lineages (Table S2). All in all, this implies that >90% of known oncogenes with autoinhibitory folds have retained at least one ohnolog pair in the human genome (as well as, possibly, a few additional duplicates from more recent SSD events).

Ohnologs Are Conserved but More "Dangerous" than "Essential"

We then investigated whether the susceptibility of ohnologs to deleterious mutations could be directly quantified through comparative sequence analysis. We used Ka/Ks ratio estimates, which measure the proportion of nonsynonymous substitutions (Ka) to the proportion of synonymous substitutions (Ks) (Extended Results and Table S3). Ohnologs exhibit statistically lower Ka/Ks ratios, Figures 2, S2, and S3, which provides direct evidence of strong conservation, consistent with a higher susceptibility of ohnologs to deleterious mutations. Similar trends have also been reported for ohnologs specific to teleost fishes (Brunet et al., 2006) or to the more recent WGD in *Xenopus laevis* lineage (Sémon and Wolfe, 2008). Note, however, that the functional consequences of such deleterious mutations, leading either to a gain or a loss of function, cannot be directly inferred from Ka/Ks distributions. Yet, as outlined below, we found marked differences in the retention of "dangerous" ohnologs prone to dominant gain-of-function mutations and "essential" ohnologs exhibiting lethal loss-of-function or null mutations.

While autosomal-dominant disease genes exhibit a strong ohnolog retention bias (Figure 1B), 59% versus 35% (59%; 261/440; $p = 1.7 \times 10^{-27}$, χ^2 test), autosomal-recessive disease genes are not significantly enriched in ohnologs 37% versus 35% (37%; 221/598; $p = 0.24$, χ^2 test). Similarly, human orthologs of mouse genes, reported as being "essential" genes from large-scale null mutant studies in mouse, are not strongly enriched in ohnologs 56% versus 54% (56%; 1,537/2,729; $p = 3.8 \times 10^{-3}$, χ^2 test), where 54% = 3,190/5,956 is the global proportion of ohnologs among the 5,956 genes tested for null mutation in mouse (Experimental Procedures). In fact, this small enrichment becomes even nonsignificant once genes with dominant allelic mutants are removed from the list of 5,956 genes tested for essentiality in mouse, i.e., 50% versus 48% (50%; 760/1,525; $p = 0.09$, χ^2 test), where 48% = 1,782/3,739 is the global proportion of ohnologs among the 3,739 genes tested for essentiality in mouse, after removing dominant disease genes, oncogenes, and genes with dominant negative mutations or autoinhibitory folds.

All in all, this shows that the retention of ohnologs has been most enhanced for genes prone to autosomal-dominant deleterious mutations and not autosomal-recessive deleterious mutations. This suggests that the retention of ohnologs is more strongly related to their "dangerousness," as defined by their high susceptibility to dominant deleterious mutations, than their functional importance or "essentiality," as identified through large-scale null mutation studies in mouse.

Ultimately, we will argue that the "dangerousness" of ohnologs effectively controls their individual retention in the genomes of post-WGD species, as will be shown below in the section Model for the Retention of Dangerous Ohnologs.

Mixed Susceptibility of Human Ohnologs to Dosage Balance

An alternative hypothesis, focusing instead on the collective retention of interacting ohnologs, has been frequently invoked to account for the biased retention of ohnologs in unicellular organisms like yeast (Papp et al., 2003) or the paramecium (Aury et al., 2006) and in higher eukaryotes (Birchler et al., 2001; Makino and McLysaght, 2010).

This "dosage balance" hypothesis posits that interacting protein partners tend to maintain balanced expression levels in the course of evolution, in particular, for protein subunits of conserved complexes (Birchler et al., 2001; Veitia, 2002; Papp et al., 2003; Veitia, 2010; Makino and McLysaght, 2010). Thus, SSD of dosage balanced genes are thought to be generally detrimental through the dosage imbalance they induce, thereby raising the odds for their rapid nonfunctionalization (Papp et al., 2003; Maere et al., 2005). By contrast, rapid nonfunctionalization of ohnologs after WGD has been suggested to be opposed by dosage effect, in particular, for highly expressed genes and genes involved in protein complexes or metabolic pathways (Aury et al., 2006; Evlampiev and Isambert, 2007; Gout et al., 2010; Makino and McLysaght, 2010). This is because WGD initially preserves correct relative dosage between

Table 1. Ohnolog Families with Both Autoinhibitory and Oncogenic Properties

A. Ohnolog Receptor Tyrosine Kinases and Other Receptor Kinases								
ALK	*LTK*				*KIT*	*CSF1R*	*FLT3*	
EGFR	*ERBB2*	*ERBB3*	*ERBB4*		*MET*	*MST1R*		
FGFR1	*FGFR2*	*FGFR3*	*FGFR4*		*NPRA*	*NPRB*		
IGF1R	*INSR*	*INSRR*			*PDGFRA*	*PDGFRB*		
B. Ohnolog Cytoplasmic and Nuclear Protein Kinases								
ABL1	*ABL2*				*PKN1*	*PKN2*	*PKN3*	
ARAF	*BRAF*	*RAF1*			*PRKAA1*	*PRKAA2*		
AKT1	*AKT2*	*AKT3*			*PRKCA*	*PRKCB*	*PRKCG*	
CAMK1	*CAMK1D*	*CAMK1G*	*PNCK*		*PRKCE*	*PRKCH*		
CAMKK1	*CAMKK2*				*PRKCI*	*PRKCZ*		
CSNK1D	*CSNK1E*				*PRKD1*	*PRKD2*	*PRKD3*	
GSK3A	*GSK3B*				*PRKG1*	*PRKG2*		
GRK4	*GRK5*	*GRK6*			*PTK2*	*PTK2B*		
JAK1	*JAK2*	*JAK3*	*TYK2*		*RSK1*	*RSK2*	*RSK3*	*RSK4*
SRC	*FGR*	*FYN*	*YES1*		*MSK1*	*MSK2*		
HCK	*LCK*	*BLK*	*LYN*		*NDR1*	*NDR2*		
MKNK1	*MKNK2*				*SYK*	*ZAP70*		
NEK6	*NEK7*							
C. Ohnolog GEF								
ARHGEF3	*NET1*				*RALGDS*	*RGL1*	*RGL2*	*RGL3*
ARHGEF6	*COOL1*				*SOS1*	*SOS2*		
DBL	*DBS*	*MCF2L2*			*TIAM1*	*TIAM2*		
FGD1	*FGD2*	*FGD3*	*FGD4*		*TIM*	*WGEF*	*SGEF*	*NGEF*
PDZ-RHOGEF	*LSC*	*LARG*			*VAV1*	*VAV2*	*VAV3*	
P114-RHOGEF	*GEF-H1*							
D. Ohnolog GAP								
ASAP1	*ASAP2*	*ASAP3*			*PLXNA1*	*PLXNA2*	*PLXNA3*	*PLXNA4*
IQGAP1	*IQGAP2*	*IQGAP3*			*PLXNB1*	*PLXNB2*	*PLXNB3*	*PLXND1*
E. Ohnolog DNA Binding and Transcription Factors								
CEBPA	*CEBPB*	*CEBPE*			*IRF4*	*IRF8*	*IRF9*	
CUX1	*CUX2*				*MEIS1*	*MEIS2*	*MEIS3*	
ELK1	*ELK3*	*ELK4*			*p53*	*p63*	*p73*	
ETS1	*ETS2*				*RUNX1*	*RUNX2*	*RUNX3*	
ETV1	*ETV4*	*ETV5*			*SOX1*	*SOX2*	*SOX3*	
ETV6	*ETV7*							
F. Other Ohnolog Genes with Both Autoinhibitory and Oncogenic Properties								
ANP32A	*ANP32B*	*ANP32E*			*nNOS*	*eNOS*		
ATP2B1	*ATP2B2*	*ATP2B3*	*ATP2B4*		*NOTCH1*	*NOTCH2*	*NOTCH3*	
cIAP1|2	*XIAP*				*PLCB1*	*PLCB2*	*PLCB3*	
CCNT1	*CCNT2*				*PLCD1*	*PLCD3*	*PLCD4*	
FLNA	*FLNB*	*FLNC*			*PLCG1*	*PLCG2*		
FURIN	*PCSK4*				*PTPN1*	*PTPN2*		
KPNA2	*KPNA7*				*SMURF1*	*SMURF2*		
NEDD4	*NEDD4L*				*TRPV1|3*	*TRPV2*	*TRPV4*	*TRPV5|6*
NOXA1	*NOXA2*							
G. Nonohnolog Genes with Both Autoinhibitory and Oncogenic Properties								
CAMK4	*ELF3*	*MELK*	*MOS*	*PDPK1*	*BRK*	*PTPN11*	*RET*	*RPS6KB1*	*TTN*

GEF, guanine exchange factors; GAP, GTPase activating proteins.
See also Tables S1 and S2.

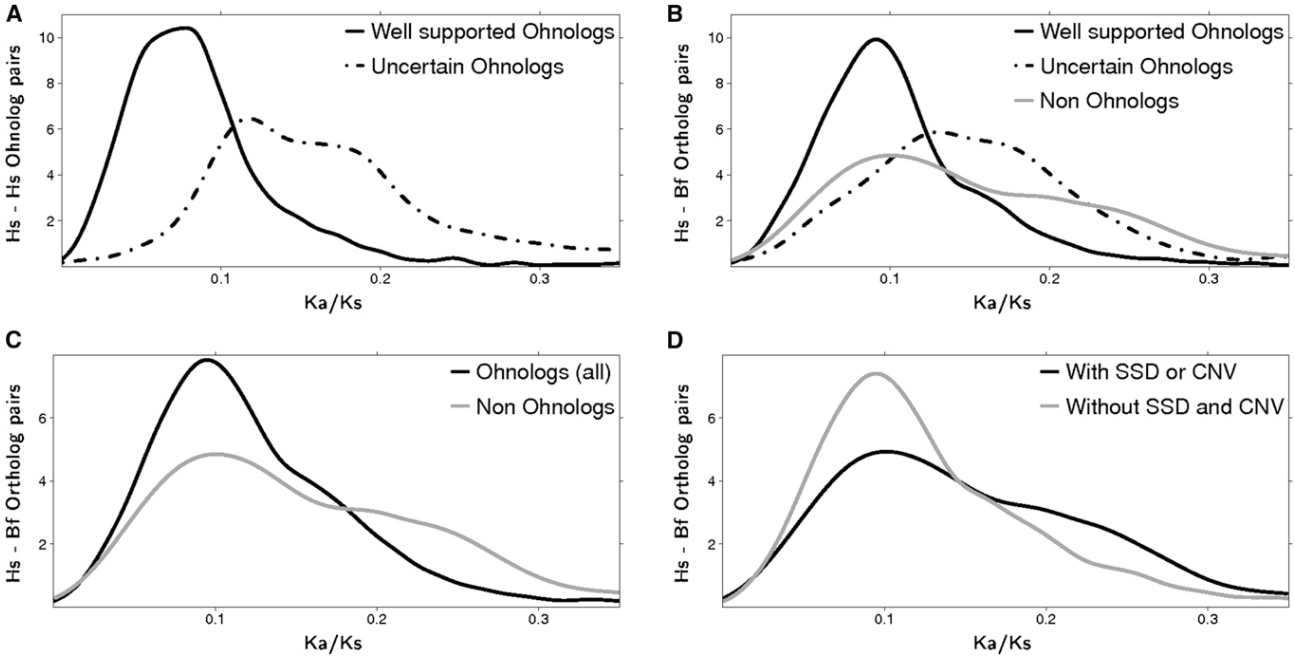

Figure 2. Ka/Ks Distributions for WGD and SSD or CNV Duplicates in the Human Genome

(A–D) Ka/Ks distributions for human-human (Hs-Hs) ohnolog pairs (A) and human-amphioxus (Hs-Bf) ortholog pairs (B) with different confidence status (see Extended Results). Ka/Ks distributions for human-amphioxus (Hs-Bf) ortholog pairs involving a human ohnolog (C) and for human-amphioxus (Hs-Bf) ortholog pairs exhibiting either SSD or CNV (D).

See also the Extended Results, Figures S2 and S3, and Table S3 for statistical significance and comparison with other invertebrate outgroups.

expressed genes, whereas subsequent random nonfunctionalization of individual ohnologs disrupts this initial dosage balance. For instance, yeast *Saccharomyces cerevisiae* has retained 76% of its ribosomal gene ohnologs from a 150 MY old WGD (Kellis et al., 2004; Lin et al., 2007), although the maintenance of these ohnologs has been suggested to require frequent gene conversion events (Kellis et al., 2004; Evangelisti and Conant, 2010) as well as fine-tuned dosage compensation to ensure a balanced expression with the remaining 24% ribosomal genes having lost their ohnologs (Zeevi et al., 2011).

Following on this dosage balance hypothesis, we performed statistical analysis on multiprotein complexes from HPRD (Keshava Prasad et al., 2009) and CORUM (Ruepp et al., 2010) databases and a hand-curated list of permanent complexes (Zanivan et al., 2007) (Experimental Procedures) to investigate for a possible association between the retention of human ohnologs and their susceptibility to dosage balance constraints.

The results depicted in Figure 1C demonstrate, in agreement with (Makino and McLysaght, 2010), that genes implicated in multiprotein complexes have retained significantly more ohnologs than expected by chance, 41% versus 35% (41%; 1,567/3,814; $p = 8.7 \times 10^{-17}$, χ^2 test). This trend is also enhanced when focusing on haploinsufficient genes, that are known for their actual sensitivity to dosage balance constraints (Qian and Zhang, 2008) (54%; 179/330; $p = 8.0 \times 10^{-14}$, χ^2 test).

Yet, surprisingly, an opposite trend corresponding to the elimination of ohnologs is observed for genes implicated in permanent complexes, that are presumably strongly sensitive to dosage balance constraints (7.5%; 18/239; $p = 1.2 \times 10^{-18}$, χ^2 test) (Figure 1C). In fact, looking more closely at the few human ohnologs, that have not been eliminated from permanent complexes (Table 2), we found that they are likely under less stringent dosage balance constraints than most proteins in permanent complexes, as they typically coassociate with mitochondrial proteins or form large multimeric subcomplexes with intrinsic stoichiometry disequilibrium.

This suggests that the elimination of most ohnologs from permanent complexes is, in fact, strongly favored under dosage imbalance and becomes likely inevitable once a few of those ohnologs have been accidentally lost following WGD. Indeed, the uneven elimination of ohnologs in permanent complexes is expected to lead to the assembly of nonfunctional, partially formed complexes detrimental to the cell, unless dosage compensation mechanisms effectively re-establish proper dosage balance at the level of gene regulation (Birchler et al., 2001), as for yeast ribosomal proteins (Zeevi et al., 2011). By contrast, transient complexes, which are typically more modular than permanent complexes, are expected to accommodate such dosage changes more easily, as they do not usually require the same strict balance in the expression levels of their protein partners.

These findings on the differences in retention of human ohnologs between permanent and more transient complexes suggest the relevance of different underlying causes. Although dosage balance presumably remains the primary evolutionary constraint in permanent complexes (<2% of human genes), which lead to the elimination of ohnologs in permanent complexes in

Table 2. Low Retention of Ohnologs in Permanent Complexes

Permanent Complexes[a]	Number of Ohnologs	Intrinsic Stoichiometry Disequilibrium of Ohnologs in Permanent Complexes
ATP F0	3/12	the 3 ohnologs ATP5G1-3 form the 10-mer C-ring of the F-type ATP synthase
ATP F1	0/5	
COX	2/11	the 2 ohnologs COX4I1,2 coassemble with 3 mitochondrial encoded genes
SRS	2/32	Ohnologs are X-linked RPS4X (with no X-inactivation) and Y-linked RPS4Y1
Mitochondrial SRS	0/30	
LRS	2/50	RPL3 and RPL39 have ohnologs RPL3L and RPL39L with unknown functions
Mitochondrial LRS	0/48	
Proteasome	2/31	ohnologs PSMA7 or PSMA7L are included in the 2 rings of 7 α subunits
Pyruvate dehydrogenase	0/5	
RNA Pol II	0/12	
RNA Pol III	0/9	

COX, cytochrome c oxidase; LRS, large ribosomal subunit; SRS, small ribosomal subunit.
[a]Zanivan et al., 2007.

vertebrate genomes, gene susceptibility to deleterious mutations may be more relevant for the retention of ohnologs within the 17% of human genes participating in more transient complexes. For instance, transient complexes involved in phosphorylation cascades or GTPase signaling pathways are known to be more sensitive to the level of activation of their protein partners than to their total expression levels. Thus, although the active forms of multistate proteins typically amount to a small fraction of their total expression level, hence providing a large dynamic range for signal transduction, it also makes them particularly susceptible to gain-of-function mutations. Such mutations can shift protein activation levels 10- to 100-fold without changes in expression levels and likely underlie stronger evolutionary constraints than the 2-fold dosage imbalance caused by gene duplication.

Indirect Cause of Ohnolog Retention in Protein Complex

To further investigate the relative effects of dosage balance and gene susceptibility to deleterious mutations, we analyzed whether the overall enhanced retention of ohnologs within multiprotein complexes (Figure 1C) could indirectly result from an enhanced susceptibility to deleterious mutations. Indeed, as outlined in Figure 3A, cancer and disease genes are more prevalent within complexes than expected by chance, 29% versus 19% (29%; 2,362/8,095; $p = 3.7 \times 10^{-132}$, χ^2 test) and this trend is enhanced for genes with stronger susceptibility to deleterious mutations, such as oncogenes (39%; 320/813; $p = 2.9 \times 10^{-52}$, χ^2 test) or oncogenes with autoinhibitory folds (59%; 67/114; $p = 2.9 \times 10^{-28}$, χ^2 test). By contrast, ohnologs are only slightly, although significantly, more prevalent in complexes than expected by chance, 22% versus 19% (22%; 1,567/7,110; $p = 9.0 \times 10^{-14}$, χ^2 test), whereas the proportion implicated in cancer or disease genes is clearly enhanced 54% versus 39% (54%; 3,844/7,110; $p = 9.5 \times 10^{-140}$, χ^2 test).

To go beyond these simple statistical associations and quantify the direct versus indirect effects of deleterious mutations and dosage balance constraints on the biased retention of human ohnologs, we have performed a Mediation analysis following the approach of Pearl (Pearl, 2001, 2011). The Mediation framework, developed in the context of causal inference analysis, aims at uncovering, beyond statistical correlations, causal pathways along which changes in multivariate properties are transmitted from a cause, *X*, to an effect, *Y*. More specifically, a Mediation analysis assesses the importance of a mediator, *M*, in transmitting the indirect effect of *X* on the response $Y \equiv Y(x, m(x))$ (Figure 3B).

Mediation analyses have been typically used in social sciences research (Baron and Kenny, 1986) as, for instance, in the context of legal disputes over alleged discriminatory hiring. In such cases, the problem is to establish that gender or race (*X*) have directly influenced hiring (*Y*) and not simply indirectly through differences in qualification or experience (*M*). Mediation analyses have also been used in epidemiology, as in a formal study (Robins and Greenland, 1992) that establishes the direct effect of smoking (*X*) on the incidence of cardiovascular diseases (*Y*), while taking into account the indirect effect of other aggravating factors, such as hyperlipidemia (*M*).

In this report, we have applied the Mediation analysis to genomic data to discriminate between direct effect (*DE*) and indirect effect (*IE*) of deleterious mutations (*X* or *M*) and dosage balance constraints (*M* or *X*) on the biased retention of human ohnologs (*Y*). The results, derived in Extended Experimental Procedures (Table S4) and summarized in Figure 3C and Table S5, demonstrate that the retention of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multiprotein complexes.

Indeed, the direct causal effect of a change from "noncomplex" to "complex" proteins only accounts for 23% of a small total effect (*TE*) of complex on the retention of ohnologs (*DE*/*TE* = 23% with *TE* = 0.079), whereas 82% of this small total effect is indirectly mediated by their susceptibility to deleterious mutations (*IE*/*TE* = 82% with 5% nonlinear coupling between direct and indirect effects) (Extended Results). By contrast, the alternative hypothesis, assuming a direct effect of deleterious mutations, accounts for 99% of a three times larger total effect on ohnolog retention (*DE*/*TE* = 99% with *TE* = 0.23), whereas the "complex" versus "noncomplex" status of human genes

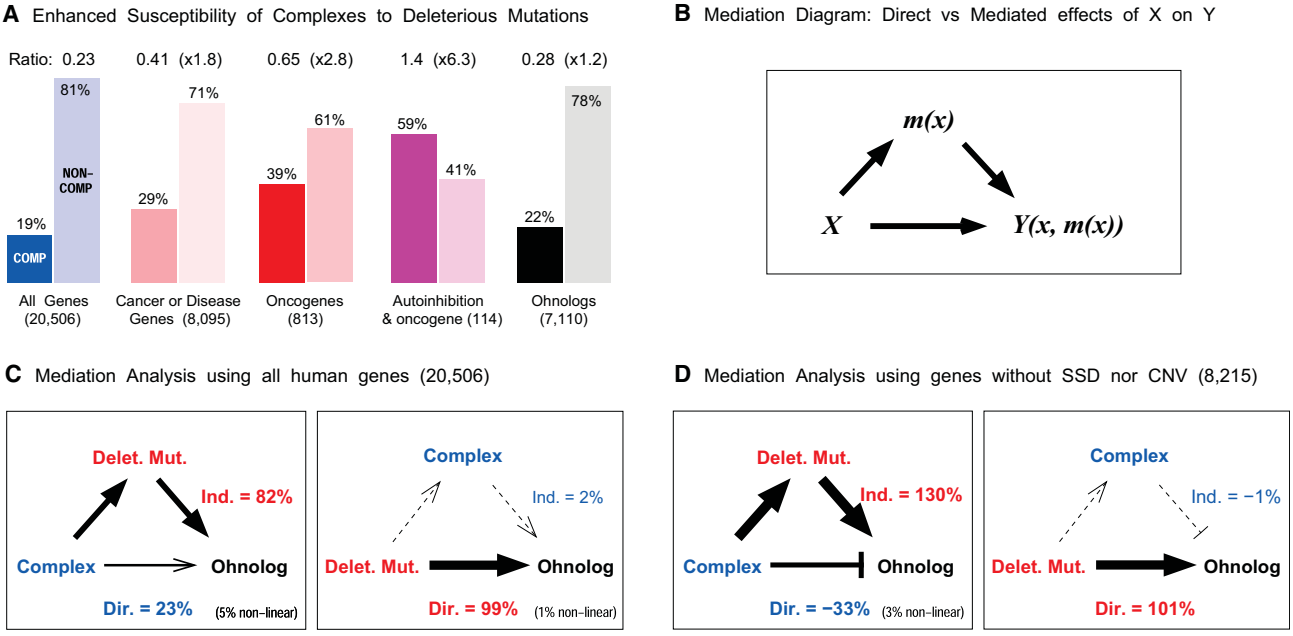

Figure 3. Mediation Analysis of the Indirect Effect of Deleterious Mutations on the Retention of Ohnologs in Multiprotein Complexes

(A) Enhanced susceptibility of complexes to deleterious mutations.

(B) Mediation diagram depicting the direct versus indirect (i.e., mediated) effects of the cause *X* on the outcome *Y*(*x*,*m*(*x*)) (Pearl, 2011). See also Extended Experimental Procedures.

(C and D) Quantitative Mediation analysis of direct versus indirect effects of deleterious mutations and dosage balance on the retention of human ohnologs using (C) all human genes (20,506) or (D) all human genes without SSD nor CNV (8,215). The thickness of the arrows outlines the relative importance of the corresponding direct or indirect effects. These results are consistent with those obtained from partial correlation analysis.

See also the main text, Extended Results, and Tables S4, S5, and S6.

has a negligible indirect effect on ohnolog retention in this case (*IE*/*TE* = 2%) (Extended Results). These trends are also further enhanced when the analysis is restricted to the 40% of human genes (8,215) without SSD and CNV duplicates (Figure 3D; Table S5; Extended Results). In fact, the direct effect of multiprotein complexes then tends to oppose the retention of ohnologs (*DE*/*TE* = −33% with *TE* = 0.064), as in the case of permanent complexes detailed above, but on an increased sample size of 8,215 genes without SSD or CNV duplicates (i.e., more than a third of human genes) in place of 239 genes from permanent complexes. By contrast, there is a five times larger total effect due to the direct effect of deleterious mutations on the retention of ohnologs (*DE*/*TE* = 101% with *TE* = 0.32), Figure 3D. This is an instance of Simpson's paradox, where two effects oppose each other, thereby, revealing the existence of conflicting underlying causes, namely, a strong positive effect of deleterious mutations and a small negative effect of dosage balance constraints on the retention of human ohnologs without SSD and CNV duplicates.

We have also examined the effects of other alternative properties on the retention of ohnologs (Extended Results; Table S5). In particular, we have found that gene expression levels and Ka/Ks ratios do not significantly mediate the effect of deleterious mutations on the retention of ohnologs. In fact, gene expression levels (Extended Experimental Procedures) have a negligible total effect on the retention of human ohnologs (*TE* = 0.003), by contrast to what has been reported for the paramecium (Gout et al., 2009). The total effects of Ka/Ks on ohnolog retention are also lower than the total effects of deleterious mutations, as *TEs* from deleterious mutations are ~2- to 3-fold stronger than *TEs* from Ka/Ks and become >10-fold stronger for genes without SSD and CNV (Extended Results).

In addition, we have performed a complementary systematic study of all these genomics properties using partial correlation analysis, which aims at "removing" the effect of a third property (*Z*) on the standard pair correlations between two variables (*X*) and (*Y*). The results detailed in Extended Results and Table S6 are entirely consistent with those obtained through mediation analysis, although the two approaches are not equivalent. Indeed, although mediation effects require partial correlation, partial correlation does not imply mediation, in general (Extended Results).

All in all, these results support the fact that the retention of ohnologs in the human genome is more strongly associated with their "dangerousness" (i.e., susceptibility to dominant deleterious mutations) than with their functional importance ("essentiality"), sensitivity to dosage balance, absolute expression levels or sequence conservation (i.e., Ka/Ks).

Model for the Retention of "Dangerous" Ohnologs

As demonstrated above, human genes with a documented sensitivity to dominant deleterious mutations have retained statistically more ohnologs from the two WGD events at the

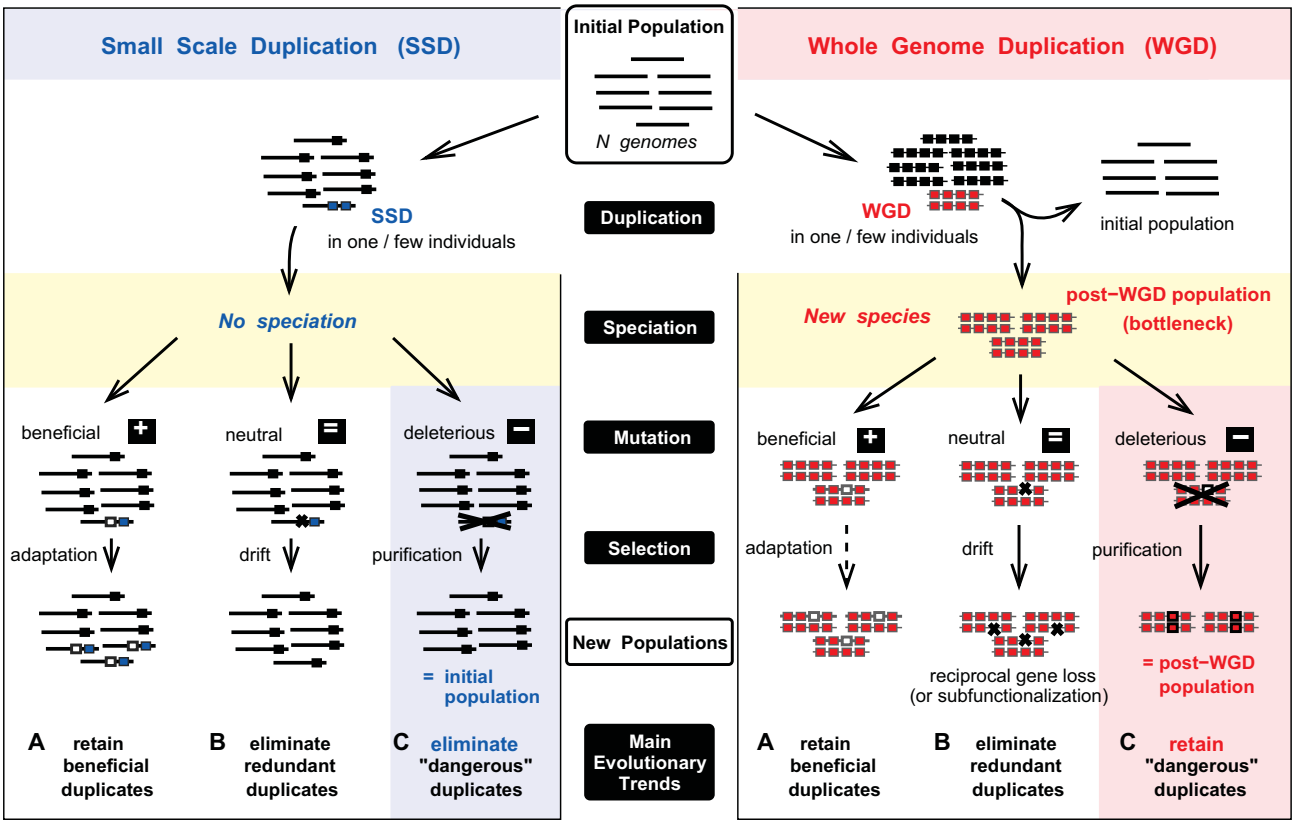

Figure 4. Evolutionary Trends of Duplicated Genes following SSD or WGD

(A–C) Horizontal lines represent the genome of different individuals. Square blocks symbolize the genes, duplicated (SSD: blue; WGD: red) or not (black). Black crosses highlight the loss of one gene (small crosses) or the elimination of an individual (larger crosses), whereas bordered square blocks emphasize retained mutated copies. Evolutionary scenarios are depicted at the population genetics level following either a SSD (left panel) or a WGD (right panel) in one or a few individuals of an initial population. Unlike SSD, WGD is invariably coupled to a speciation event, owing to the difference in ploidy between pre- and post-WGD individuals. Three possible scenarios—beneficial (A), neutral or nearly neutral (B), or deleterious mutations (C) in one gene duplicate—are outlined in post-SSD and post-WGD populations. The main difference concerns the mutation/selection process of "dangerous" genes, i.e. genes prone to autosomal-dominant deleterious mutations (C). See main text for a detailed description.

onset of jawed vertebrates. This suggests that ohnologs have been retained in vertebrate genomes, not because they initially brought selective advantages following WGD, but because their mutations were more likely detrimental or lethal than nonfunctional, thereby preventing their rapid elimination from the genomes of surviving individuals following WGD transitions, as outlined in the evolutionary model depicted in Figure 4.

For completeness and clarity, Figure 4 examines all possible evolutionary scenarios following either a SSD or a WGD duplication event in the genome of one or a few individuals in an initial population. The first and critical difference between SSD and WGD duplication events occurs at the population genetics level with an obligate speciation following WGD event, owing to the difference in ploidy between pre- and post-WGD individuals. As a result, all individuals in the post-WGD population carry twice as many genes as their pre-WGD relatives, whereas only a few individuals in the post-SSD population carry a single small duplicated region. Figure 4 then outlines the three mutation/selection scenarios focusing on a single gene duplicate in the genomes of post-SSD or post-WGD populations: (A) Beneficial mutations after SSD or WGD are expected to spread and become eventually fixed in the new populations, although the bottleneck in population size following WGD limits in practice the efficacy of adaptation in post-WGD species. (B) Neutral or nearly neutral mutations mainly lead to the random nonfunctionalization of one copy of most redundant gene duplicates and, therefore, to their elimination following both SSD and WGD events. In post-WGD populations, this results in the "reciprocal gene loss" of most gene duplicates, which is also known to lead to further speciations in post-WGD species, owing to the interbreeding incompatibility between post-WGD individuals with different "reciprocal gene loss" pattern (Lynch and Force, 2000a). Alternatively, neutral or nearly neutral mutations can also result in the eventual retention of both duplicate copies through subfunctionalization (Hughes, 1994; Lynch and Force, 2000b), that is, by rendering each duplicate copy unable to perform all the functions of their ancestral gene (see Discussion). (C) Finally, dominant deleterious mutations favor the elimination of the individuals

(or their descendants) harboring them through purifying selection. However, this typically leads to opposite outcomes in post-SSD and post-WGD populations. In post-SSD populations, dominant deleterious mutations will tend to eliminate SSD duplicates before they have the time to reach fixation (see below). By contrast, in post-WGD populations, where all ohnologs have been initially fixed through WGD-induced speciation, purifying selection will effectively favor the retention of dangerous ohnologs, as all surviving individuals still present (nondeleterious) functional copies of these dangerous genes.

Note, in particular, that this somewhat counterintuitive evolutionary model for the retention of "dangerous" ohnologs hinges on two unique features:

(1) It requires an autosomal dominance of deleterious mutations, in agreement with our observation, above, that retained ohnologs are more "dangerous" than "essential."
(2) It relies on the fact that successful WGD events start with a concomitant speciation event, which immediately fixes all ohnolog duplicates in the initial post-WGD population (Figure 4).

Note, also, that the same evolutionary trend is expected for dangerous SSD duplicates that would have the time (t) to become fixed through genetic drift in a population of size N before deleterious mutations can arise at a rate K, i.e., $t = 4N < 1/K$. This corresponds to a population bottleneck effect with $N < 1/(4K) \approx 5{,}000–10{,}000$ for typical vertebrates.

DISCUSSION

Beyond human and vertebrate genomes, WGD events have now been established in all major eukaryote kingdoms (Sémon and Wolfe, 2007; Evlampiev and Isambert, 2007). Unlike SSD events, WGD transitions provide a unique evolutionary mechanism, enabling the simultaneous duplication of entire genetic pathways and multiprotein complexes, followed by long periods of functional divergence and extensive loss of ohnologs (Aury et al., 2006). Moreover, although both WGD and SSD events have expanded the gene repertoires and resulting protein networks (Evlampiev and Isambert, 2007; Evlampiev and Isambert, 2008) of eukaryotes, it has become increasingly clear that WGD and SSD events actually lead to the expansion of different gene classes in the course of evolution, (Maere et al., 2005; Aury et al., 2006; Sémon and Wolfe, 2007; Makino and McLysaght, 2010; Huminiecki and Heldin, 2010; and this study).

In this article, we report that WGD have effectively favored the expansion of gene families prone to deleterious mutations in the human genome, such as genes implicated in cancer and genes with autoinhibitory interactions. In particular, we found that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations.

From a broader perspective, a number of studies have now shown that many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent. In the light of the present study, we expect that many of these statistically significant correlations mainly result from indirect rather than direct associations, which may even frequently oppose each other. This highlights the need to rely on more advanced inference methods to analyze the multiple, direct, and indirect causes underlying the evolution of specific gene repertoires.

In the present study, we have quantitatively analyzed the direct versus indirect effects of the susceptibility of human genes to deleterious mutation and dosage balance constraints on the retention of ohnologs and proposed a simple evolutionary mechanism to account for the initial retention of "dangerous" ohnologs after WGD (Figure 4). On longer timescales, we expect that this initial retention bias of "dangerous" ohnologs effectively promote a prolonged genetic drift and, thus, a progressive functional divergence between ohnolog pairs. This eventually favors the subfunctionalization (Hughes, 1994; Lynch and Force, 2000b) of ancestral functions between ohnolog pairs, which ultimately warrants their long-term maintenance following WGD events.

Note, however, that this subfunctionalization process requires that the expression of ohnologs is not rapidly suppressed by large-scale deletion or silencing mutations in regulatory regions. As ohnolog pairs are not arranged in tandem, large-scale deletions through unequal crossing-over cannot typically remove entire ohnolog duplicates while preserving the integrity of nearby genes. Furthermore, as the size of promoter or enhancer regions is typically much smaller than UTRs and coding regions, one expects that the rate of transcriptional silencing does not exceed the rates of functional silencing and divergence in UTRs and coding regions. In fact, early estimates (Nadeau and Sankoff, 1997) showed that gene loss and functional divergence after genome duplications in early vertebrates occurred at comparable rates in gene families including at least two ohnologs. This is also directly evidenced by pseudotetraploid species like the vertebrate *Xenopus laevis*, which still retains ~40% of its initial ohnologs from a 30-million-year-old WGD (Sémon and Wolfe, 2008). All in all, this suggests that ohnologs prone to dominant deleterious mutations have at least a few million years to diverge and become nonredundant genes before they have a chance to be deleted or transcriptionally silenced.

Yet, we found that the retention of these dangerous ohnologs remains intrinsically stochastic by nature as many of them have also been eliminated following WGD events. This presumably occurred through loss-of-function mutations, transcriptional silencing, or large-scale deletion before ohnolog pairs could diverge and become nonredundant genes. More quantitatively, a simple theoretical estimate, derived from the long-term retention statistics of Figure 1, shows that only 6%–10% of the initial ohnolog duplicates have been retained on average at each round of WGD, Figure 5 (see Extended Results for details). By comparison, ~23%–30% of the initial ohnologs prone to gain-of-function mutations have been retained on average at each WGD. This implies that genes susceptible to deleterious mutations are two to five times more likely to retain ohnologs on long evolutionary timescales. Moreover, genes combining several factors associated with enhanced susceptibility to autosomal-dominant deleterious mutations are shown to be more than ten times more

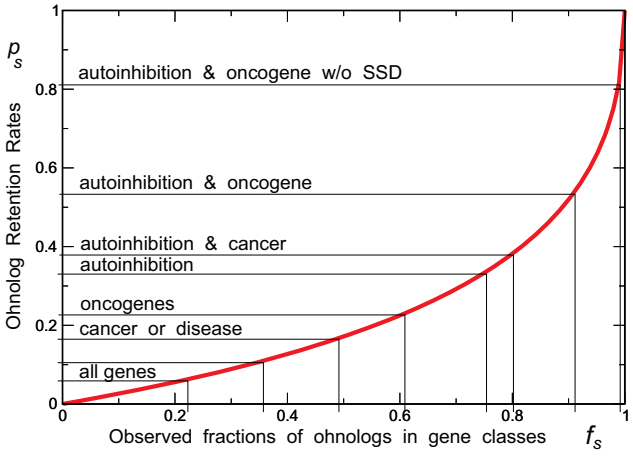

Figure 5. Estimates of Ohnolog Retention Rates

Estimates of ohnolog retention rates p_s in early vertebrates from the observed fraction f_s of ohnologs in the human genome for gene classes, s, with increasing susceptibility to deleterious mutations. The theoretical estimate (red curve) is obtained assuming that the retentions of ohnologs were comparable for each of the two WGD at the onset of vertebrates, and reads $P_s = 2/f_s - 1 - \sqrt{(2/f_s - 1)^2 - 1}$ as detailed in the Extended Results and Tables S7 and S8.

likely to retain ohnologs than genes lacking gain-of-function mutations (Figure 5), as illustrated on the examples of oncogenes with autoinhibitory folds (Table 1).

In turn, the elimination of ohnologs has been shown to drive further speciation events within post-WGD (sub)populations, due to the emergence of recombination barriers from the accumulation of differences in ohnolog deletion patterns between post-WGD individuals (Lynch and Force, 2000a). The resulting fragmentation of post-WGD subpopulations is then expected to sustain negative selection pressure that favors the retention of the remaining ohnolog pairs prone to deleterious mutations, as outlined in Figure 4. Hence, although most WGDs are unlikely to bring much fitness benefit on short evolutionary timescales (if only due to the population bottlenecks associated with WGD-induced speciations; Figure 4), they provide a unique evolutionary mechanism to experiment virtually unlimited combinations of regulation/deletion patterns from redundant ohnolog genes. Over long timescales (>100–500 MY), such trial and error combinations have visibly led to the evolutionary success and radiation of WGD species.

In summary, we present evidence supporting an evolutionary link between the susceptibility of human genes to dominant deleterious mutations and the documented expansion of these "dangerous" gene families by two WGD events at the onset of jawed vertebrates. We propose that deleterious mutations, responsible for many cancers and other severe genetic diseases on the lifespan of human individuals, have also underlain purifying selection over long evolutionary timescales, which effectively favored the retention of vertebrate ohnologs prone to dominant deleterious mutations, as outlined in Figure 4. From a population genetics perspective, we argue that this counterintuitive retention of dangerous ohnologs hinges in fact on WGD-induced speciation events, which are largely credited for the genetic complexity and successful radiation of vertebrate species.

These findings highlight the importance of purifying selection from WGD events on the evolution of vertebrates and, beyond, exemplify the role of nonadaptive forces on the emergence of eukaryote complexity (Fernández and Lynch, 2011).

EXPERIMENTAL PROCEDURES

WGD Duplicated Genes or "Ohnologs"

Human ohnolog genes were obtained from (Makino and McLysaght, 2010). Makino and McLysaght compared different vertebrate and six nonvertebrate outgroup genomes to identify ohnologs in the human genome. The final data set consists of 8,653 ohnolog pairs and 7,110 unique ohnologs. We further divided ohnologs into well supported (3,963), plausible (894), and more uncertain (2,253) ohnologs (see Extended Experimental Procedures).

SSD Duplicated Genes

We identified paralogous genes within the human genome from sequence similarity search. We obtained a total of 11,185 SSD genes. In particular, paralogs that were not annotated as ohnologs were taken to be SSD genes (see Extended Experimental Procedures).

Genes with CNV

CNV regions were obtained from Database of Genomic Variants (Zhang et al., 2006). A total of 5,709 genes were identified to be CNV genes as their entire coding sequence fell within one of the CNV regions.

Cancer and Disease Genes

We obtained cancer genes from multiple databases, including COSMIC (Forbes et al., 2011) and CancerGenes (Higgins et al., 2007), listed in Table S7. The detailed list of 6,917 cancer genes is given in Table S8 with a hand-curated list of 813 verified or predicted (Bozic et al., 2010) oncogenes (see Extended Experimental Procedures). We obtained 2,580 disease genes from the "Morbidmap" database of OMIM and hand curated subsets of 440 autosomal-dominant and 598 autosomal-recessive disease genes from Blekhman et al. (2008).

Genes with Autoinhibitory Folds

To obtain genes coding for proteins with autoinhibitory folds we searched PubMed with keyword "autoinhibitory domain" and retrieved relevant autoinhibitory genes and domains manually. Further gene candidates with autoinhibitory folds were obtained from databases, OMIM, SwissProt, NCBI Gene, and GeneCards using the parsing terms: auto/self-inhibit*. Careful manual curation of this list of gene candidates with the available literature finally yielded a total of 461 genes with autoinhibitory folds (94% of initial candidates).

Essential Genes

Mouse essential genes were obtained from Mouse Genome Informatics database. Essential genes were defined as genes having lethal or infertility phenotypes on loss-of-function or knockout mutations (2,729 genes) (see Extended Experimental Procedures).

Genes in Complexes and Permanent Complexes

Protein complexes were obtained from Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009) and CORUM database (Ruepp et al., 2010). In addition, a manually curated data set of permanent complexes (239 genes) was obtained from Zanivan et al. (2007). The final data set consists of 3,814 protein complex genes (see Extended Experimental Procedures).

Haploinsufficient and Dominant Negative Genes

Haploinsufficient and dominant negative candidate genes were obtained from parsing OMIM text files with Perl regular expressions. The resulting gene lists were manually curated with the available literature, yielding a total of

330 haploinsufficient genes (80% of initial candidates) and 477 dominant-negative genes (63% of initial candidates).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Results, Extended Experimental Procedures, three figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2012.09.034.

LICENSING INFORMATION

ACKNOWLEDGMENTS

P.P.S. acknowledges a PhD fellowship from Erasmus Mundus and Université Pierre et Marie Curie; S.A. acknowledges a PhD fellowship from Ministry of Higher Education and Research, France; I.C. acknowledges postdoctoral support from ANR (grant ANR-08-BLAN-0290); R.S. acknowledges PhD fellowships from INCa and ARC; H.I. and J.C. acknowledge funding from Foundation Pierre-Gilles de Gennes. We thank H. Roest Crollius, L. Peliti, S. Coscoy and V. Hakim for discussions.

Received: April 12, 2012
Revised: September 17, 2012
Accepted: September 27, 2012
Published: November 15, 2012

REFERENCES

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. *51*, 1173–1182.

Berry, L.W., Westlund, B., and Schedl, T. (1997). Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors. Development *124*, 925–936.

Birchler, J.A., Bhadra, U., Bhadra, M.P., and Auger, D.L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev. Biol. *234*, 275–288.

Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. *7*, R43.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., and Nowak, M.A. (2010). Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. USA *107*, 18545–18550.

Brunet, F.G., Roest Crollius, H., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. *23*, 1808–1816.

Cai, J.J., Borenstein, E., Chen, R., and Petrov, D.A. (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol. Evol. *1*, 131–144.

Ciocan, C.M., Moore, J.D., and Rotchell, J.M. (2006). The role of ras gene in the development of haemic neoplasia in Mytilus trossulus. Mar. Environ. Res. Suppl. *62*, S147–S150.

Dickerson, J.E., and Robertson, D.L. (2012). On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol. Biol. Evol. *29*, 61–69.

Domazet-Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol. Biol. Evol. *25*, 2699–2707.

Esteban, L.M., Vicario-Abejón, C., Fernández-Salguero, P., Fernández-Medarde, A., Swaminathan, N., Yienger, K., Lopez, E., Malumbres, M., McKay, R., Ward, J.M., et al. (2001). Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development. Mol. Cell. Biol. *21*, 1444–1452.

Evangelisti, A.M., and Conant, G.C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. Genome Biol. Evol. *2*, 826–834.

Evlampiev, K., and Isambert, H. (2007). Modeling protein network evolution under genome duplication and domain shuffling. BMC Syst. Biol. *1*, 49.

Evlampiev, K., and Isambert, H. (2008). Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc. Natl. Acad. Sci. USA *105*, 9863–9868.

Fernández, A., and Lynch, M. (2011). Non-adaptive origins of interactome complexity. Nature *474*, 502–505.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*(Database issue), D945–D950.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. *16*, 805–814.

Furney, S.J., Albà, M.M., and López-Bigas, N. (2006). Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics *7*, 165.

Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. *14*, 46–49, discussion 49–50.

Gout, J.F., Duret, L., and Kahn, D. (2009). Differential retention of metabolic genes following whole-genome duplication. Mol. Biol. Evol. *26*, 1067–1072.

Gout, J.F., Kahn, D., and Duret, L.; Paramecium Post-Genomics Consortium. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. *6*, e1000944.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics *175*, 933–943.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. *8*, R209.

Higgins, M.E., Claremont, M., Major, J.E., Sander, C., and Lash, A.E. (2007). CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res. *35*(Database issue), D721–D726.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. *256*, 119–124.

Huminiecki, L., and Heldin, C.H. (2010). 2R and remodeling of vertebrate signal transduction engine. BMC Biol. *8*, 146.

Ise, K., Nakamura, K., Nakao, K., Shimizu, S., Harada, H., Ichise, T., Miyoshi, J., Gondo, Y., Ishikawa, T., Aiba, A., and Katsuki, M. (2000). Targeted deletion of the H-ras gene decreases tumor formation in mouse skin carcinogenesis. Oncogene *19*, 2951–2956.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37*(Database issue), D767–D772.

Lin, Y.S., Hwang, J.K., and Li, W.H. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome. Gene *387*, 109–117.

Lynch, M., and Force, A. (2000a). Gene duplication and the origin of interspecific genomic incompatibility. Am. Nat. *156*, 590–605.

Lynch, M., and Force, A. (2000b). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459–473.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA *102*, 5454–5459.

Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. USA *107*, 9270–9274.

Nadeau, J.H., and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics *147*, 1259–1266.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194–197.

Pearl, J. (2001). Direct and indirect effects. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann, 411–420.

Pearl, J. (2011). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. In Causality: Statistical Perspectives and Applications, C. Berzuini, P. Dawid, and L. Bernardinelli, eds. (United Kingdom: John Wiley & Sons), pp. 151–175.

Pufall, M.A., and Graves, B.J. (2002). Autoinhibitory domains: modular effectors of cellular regulation. Annu. Rev. Cell Dev. Biol. *18*, 421–462.

Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064–1071.

Qian, W., and Zhang, J. (2008). Gene dosage and gene duplicability. Genetics *179*, 2319–2324.

Robert, J. (2010). Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates. Dev. Comp. Immunol. *34*, 915–925.

Robins, J.M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology *3*, 143–155.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, 497–501.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. *17*, 505–512.

Sémon, M., and Wolfe, K.H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc. Natl. Acad. Sci. USA *105*, 8333–8338.

Veitia, R.A. (2002). Exploring the etiology of haploinsufficiency. Bioessays *24*, 175–184.

Veitia, R.A. (2010). A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J. *24*, 994–1002.

Wolfe, K. (2000). Robustness—it's not where you think it is. Nat. Genet. *25*, 3–4.

Zanivan, S., Cascone, I., Peyron, C., Molineris, I., Marchio, S., Caselle, M., and Bussolino, F. (2007). A new computational approach to analyze human protein complexes and predict novel protein interactions. Genome Biol. *8*, R256.

Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011). Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. Genome Res. *21*, 2114–2128.

Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., and Scherer, S.W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. *115*, 205–214.

la sensibilité des techniques d'imagerie, permettront de continuer à développer de nouveaux rapporteurs. En outre, les outils de détection et de suivi des cellules, ainsi que l'analyse des oscillations permettront de développer le même type d'approche chez d'autres vertébrés. ◊

A technical breakthrough for understanding segmentation clock dynamics and synchrony

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Pourquié O. Vertebrate segmentation: from cyclic gene networks to scoliosis. *Cell* 2011 ; 145 : 650–63.
2. Soroldoni D, Oates AC. Live transgenic reporters of the vertebrate embryo's segmentation clock. *Curr Opin Genet Dev* 2011 ; 21 : 600–5.
3. Masamizu Y, Ohtsuka T, Takashima, Y, *et al.* Real-time imaging of the somite segmentation clock: revelation of unstable oscillators in the individual presomitic mesoderm cells. *Proc Natl Acad Sci USA* 2006 ; 103 : 1313–8.
4. Aulehla A, Wiegraebe W, Baubet V, *et al.* A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation. *Nat Cell Biol* 2008 ; 10 : 186–93.
5. Takashima Y, Ohtsuka T, Gonzalez A, *et al.* Intronic delay is essential for oscillatory expression in the segmentation clock. *Proc Natl Acad Sci USA* 2011 ; 108 : 3300–5.
6. Delaune EA, François P, Shih NP, Amacher SL. Single-cell-resolution imaging of the impact of Notch signaling and mitosis on segmentation clock dynamics. *Dev Cell* 2012 ; 23 : 995–1005.
7. Gajewski M, Sieger D, Alt B, *et al.* Anterior and posterior waves of cyclic her1 gene expression are differentially regulated in the presomitic mesoderm of zebrafish. *Development* 2003 ; 130 : 4269–78.
8. Jiang YJ, Aerne BL, Smithers L, *et al.* Notch signalling and the synchronization of the somite segmentation clock. *Nature* 2000 ; 408 : 475–9.
9. Horikawa K, Ishimatsu K, Yoshimoto E, *et al.* Noise-resistant and synchronized oscillation of the segmentation clock. *Nature* 2006 ; 441 : 719–23.

NOUVELLE

Évolution et cancer

Expansion des familles de gènes dangereux par duplication du génome

Séverine Affeldt[1*], Param Priya Singh[1*], Ilaria Cascone[2], Rasim Selimoglu[2], Jacques Camonis[2], Hervé Isambert[1]

[1]CNRS-UPMC UMR168, [2]Inserm U830, Institut Curie, Centre de recherche, 26, rue d'Ulm, 75248 Paris, France.

*Contribution égale des auteurs

herve.isambert@curie.fr

> Si la conservation des gènes essentiels à la vie des organismes se conçoit intuitivement bien, à l'inverse, l'étonnante expansion des familles de gènes à l'origine des cancers ou d'autres maladies génétiques chez les vertébrés pose question. Alors qu'on pourrait supposer que la multiplication de ces gènes « dangereux » confère malgré tout un avantage sélectif, nos travaux récents [1] suggèrent en fait que ces gènes ont été multipliés et conservés en raison de leur dangerosité à la suite de deux accidents génétiques majeurs correspondant à des duplications globales de génome.

De l'expansion des familles de gènes dangereux chez les vertébrés

Pour comprendre l'origine de l'expansion des familles de gènes dangereux chez les vertébrés, il faut remonter à l'ancêtre commun de tous les vertébrés, il y a quelque 500 millions d'années. Par un mécanisme presque toujours létal, mais qui a joué un rôle essentiel au cours de l'évolution, notre lignée encore invertébrée a entièrement dupliqué son génome deux fois de suite et a survécu à ces deux accidents génétiques majeurs. Ces deux duplications globales du génome ont conduit à l'émergence et à la complexification des vertébrés dont certains gènes ont conservés jusqu'à quatre copies. Au total, un quart à un tiers de nos gènes serait directement issu de ces deux duplications de génome à l'origine des vertébrés [2]. Ces gènes sont appelés gènes ohnologues en l'honneur du généticien Susumu Ohno qui fut le premier à avancer l'hypothèse de ces duplications globales du génome chez les vertébrés [3].

Après une duplication globale du génome, les organismes perdent généralement près de 80 à 90 % des gènes dupliqués, ce qui entraîne une expansion hétérogène de leurs voies de signalisation *(Figure 1)* et de leurs réseaux de gènes [4-6]. Cependant, de façon surprenante, on constate que les copies ohnologues retenues dans le génome humain comportent un nombre élevé de gènes dangereux, c'est-à-dire présentant une forte susceptibilité aux mutations délétères dominantes comme les oncogènes notamment. Certains de ces gènes dangereux ont même gardé leurs quatre copies depuis l'origine des vertébrés ! C'est le cas par exemple des gènes *RalGEF* (*Ral guanine-nucleotide exchange factor*) *(Figure 1)* qui activent les voies Ras-Ral impliquées dans la migration et la prolifération cellulaires dans des tumeurs [7]. De même, le gène *Ras*, qu'on retrouve en un seul exemplaire chez les invertébrés comme la drosophile *(Figure 1)*, a conservé chez la plupart des vertébrés trois ohnologues proto-oncogéniques (*KRas*, *HRas* et *NRas*) *(Figure 1)* qui présentent des mutations constitutivement actives dans plus de 25 % des cas de cancer chez l'homme.

DOI : 10.1051/medsci/2013294008

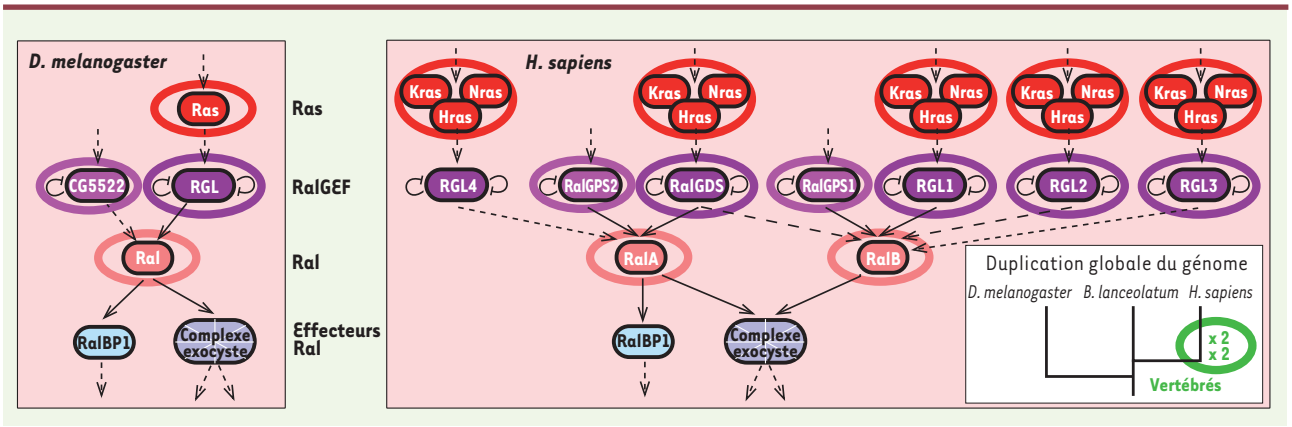

Figure 1. ***Expansion des voies de signalisation Ras-Ral.*** Voies de signalisation Ras-Ral chez la drosophile (***A***) et chez l'homme (***B***). Après les deux duplications globales de génomes chez l'ancêtre des vertébrés, chez l'homme, le gène *RGL* a conservé ses quatre copies ohnologues (*RalGDS*, *RGL1*, *RGL2*, *RGL3*), le gène *Ras* a conservé trois copies ohnologues (*Kras*, *Hras*, *Nras*) et le gène *Ral* a conservé deux copies ohnologues (*RalA*, *RalB*) [7]. Cet exemple illustre la nécessaire réorganisation des voies de signalisation et des réseaux d'interactions protéine-protéine après un évènement de duplication globale du génome, et l'élimination par divergence d'une partie des gènes dupliqués [4-6].

Au delà de ces exemples, nous avons effectué une étude d'exploration de données à grande échelle sur l'ensemble des gènes humains, à partir des bases de données accessibles en ligne (comme COSMIC [8], CancerGenes [9]) et de publications. Ceci nous a permis de mettre en évidence une forte association entre la dangerosité des ohnologues et leur rétention dans le génome humain. Dans ces analyses, nous avons considéré différentes classes de gènes susceptibles d'être affectés par des mutations délétères. Ces classes comprennent notamment des gènes dont l'implication dans les cancers est connue et des gènes dont les mutations induisent typiquement des phénotypes délétères dominants, par exemple *via* une perte d'interaction d'auto-inhibition entraînant un gain de fonction permanent pour le gène mutant.

Les résultats obtenus [1] indiquent que les 8 095 gènes impliqués dans des cancers ou dans des maladies génétiques comportent significativement plus d'ohnologues que l'ensemble des 20 506 gènes codant pour les protéines humaines, soit 48 % contre 35 % (48 % ; 3 844/8 095 ; $p = 1{,}3 \times 10^{-129}$, test χ^2). En outre, ce biais de rétention augmente fortement dans le cas des oncogènes (61 % ; 493/813 ; $p = 1{,}4 \times 10^{-54}$, test χ^2) ou des gènes dont la protéine est auto-inhibée (76 % ; 350/461 ; $p = 2{,}7 \times 10^{-77}$, test χ^2). De même, lorsque l'on considère des classes de gènes combinant plusieurs facteurs de dangerosité, telles que la classe des oncogènes dont la protéine est auto-inhibée, ces biais de rétention dépassent les 90 % (91 % ; 104/114 ; $p = 6{,}9 \times 10^{-37}$, test χ^2). À l'inverse, nous avons montré que les gènes associés à des mutations délétères récessives chez l'homme, ainsi que la plupart des gènes essentiels connus chez la souris, n'ont pas conservé un excès d'ohnologues. Il apparaît donc que la rétention des gènes ohnologues serait directement liée à leur susceptibilité aux mutations délétères dominantes et non pas à leur nature fondamentale pour l'organisme. Nous avons en fait démontré [1], au moyen d'analyses statistiques d'inférences bayésiennes, telles que l'analyse de Médiation [10], que la dangerosité est bien la cause principale de la rétention des gènes ohnologues chez l'homme. En particulier, les équilibres entre les niveaux d'expression génétique, fréquemment proposés comme la principale cause des biais de rétention des ohnologues, semblent en fait résulter indirectement des effets des mutations délétères dominantes [1].

Le mécanisme de conservation des ohnologues

Pourquoi les multiples copies ohnologues de ces gènes dangereux, à l'origine de nombreux cancers et maladies génétiques sévères, n'ont-elles pas été éliminées chez les vertébrés ? Pour le comprendre, il faut avoir en tête deux points essentiels *(Figure 2)* : (1) la duplication globale du génome, lorsqu'elle n'est pas létale, implique nécessairement l'apparition d'une nouvelle espèce composée d'individus possèdant tous initialement leurs gènes en double ; et (2) l'inégalité des gènes face aux mutations. Alors que la plupart des gènes tendent à perdre leur fonction par mutation, les gènes dangereux se caractérisent par le fait que leurs mutations entraînent fréquemment une suractivation, c'est-à-dire un gain de fonction, plutôt qu'une perte de fonction. En général, la perte de fonction d'un ohnologue ne pose pas de problème tant qu'il reste une copie fonctionnelle de ce gène, ce qui conduit à l'élimination progressive d'une des copies de la plupart des ohnologues non dangereux. En revanche, la survenue de mutations conduisant à des gains de fonction ou, plus généralement, à des phénotypes dominants délétères qui caractérisent les ohnologues dangereux, va entraîner des pathologies du développement ou

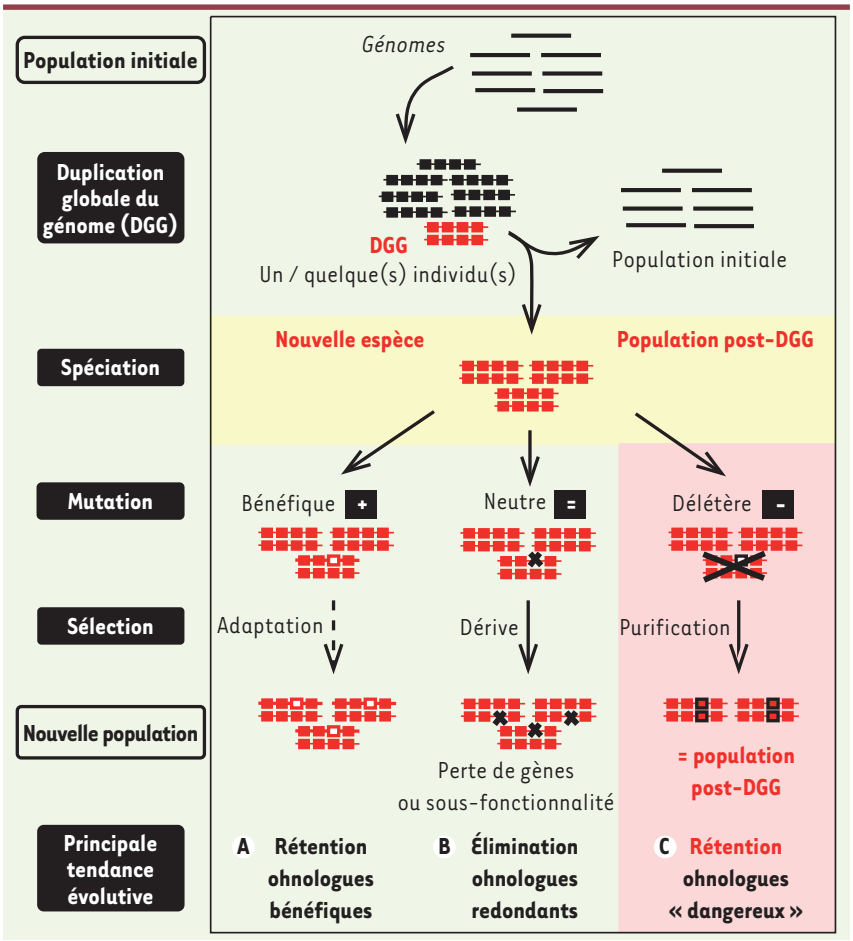

Figure 2. ***Évolution des gènes dupliqués après duplicaton globale du génome. A-C.*** Les lignes horizontales représentent le génome de différents individus. Les carrés symbolisent les gènes dupliqués (rouges) ou non dupliqués (noirs). Les croix noires représentent la perte d'un seul gène (petites croix) ou l'élimination d'un individu (grande croix), tandis que les carrés avec bordures indiquent les copies mutées. Les scénarios d'évolution sont décrits comme il se doit au niveau d'une population d'individus. Lorsqu'elle n'est pas létale, une duplication globale de génome chez un ou plusieurs individus de la population initiale (deux lignes horizontales rouges pour un individu avec duplication complète) implique nécessairement l'émergence d'une nouvelle espèce dont tous les gènes ont été dupliqués. La sélection de purification favorise alors indirectement la conservation des ohnologues dangereux non mutés dans le génome des individus qui survivent. Adapté de [1].

des tumeurs. Celles-ci pénalisent les organismes atteints et, plus ou moins directement, leur descendance qui finira par s'interrompre. Pour autant, les gènes dangereux impliqués ne sont pas éliminés mais au contraire conservés dans la population, puisqu'ils sont encore présents sous une forme non délétère dans le reste de la population issue de la duplication du génome *(Figure 2)*. Ce processus évolutif par élimination de mutants (sélection de purification) se distingue du concept d'avantage sélectif généralement associé à l'évolution (sélection naturelle ou adaptative).

La multiplication des gènes dangereux chez les vertébrés est donc liée à ce phénomène spécifique et très rare de duplication globale du génome qui existe en fait dans la plupart des branches eucaryotes. Sa compréhension doit se faire au niveau de la génétique des populations au sein des nouvelles espèces issues d'une duplication de génome. Tout se passe comme si les individus de ces populations n'avaient pas pu se débarrasser de nombreux gènes dangereux redondants, directement fixés par spéciation dans leur génome dupliqué et non pas fixés progressivement grâce à un avantage sélectif comme pour la plupart des gènes dupliqués individuellement. Ensuite, les gènes ohnologues conservés se différencient et deviennent souvent des acteurs majeurs du développement, de la signalisation et de la régulation cellulaires. Par exemple, les cadhérines, sorte de colle qui lient les cellules entre elles, ont conservé de multiples ohnologues exprimés dans différents tissus, comme la E-cadhérine qui lie les cellules épithéliales entre elles, ou la N-cadhérine exprimée dans le tube neural et les neurones. Mais les mutations des cadhérines qui entraînent la décohésion des cellules entre elles sont aussi impliquées dans la migration des cellules tumorales et leur dissémination vers d'autres organes.

Conclusions

Ces accidents génétiques majeurs de doublement du génome survenus il y a 500 millions d'années dans l'évolution des vertébrés ont donc permis l'émergence d'organismes plus complexes, mais aussi la multiplication des gènes dangereux chez les vertébrés. Ces résultats éclairent d'un point de vue évolutif nouveau l'expansion des familles de gènes fréquemment impliquées dans des maladies génétiques et de nombreux cancers. ◊

Evolution and cancer: expansion of dangerous gene repertoire by whole genome duplications

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Singh PP, Affeldt S, Cascone I, *et al*. On the expansion of dangerous gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep* 2012 ; 2 : 1387-98.
2. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA* 2010 ; 107 : 9270-74.

3. Ohno S. *Evolution by gene duplication*. New York : Springer-Verlag, 1970.
4. Evlampiev K, Isambert H. Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst Biol* 2007 ; 1 : 49.
5. Evlampiev K, Isambert H. Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc Natl Acad Sci USA* 2008 ; 105 : 9863-8.
6. Stein RR, Isambert H. Logistic map analysis of biomolecular network evolution. *Phys Rev E* 2011 ; 84 : 051904.
7. Cascone I, Selimoglu R, Ozdemir C, *et al.* Distinct roles of RalA and RalB in the progression of cytokinesis are supported by distinct RalGEFs. *EMBO J* 2008 ; 27 : 2375-87.
8. Forbes SA, Bindal N, Bamford S, *et al*. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2011 ; 39 : D945–D950.
9. Higgins ME, Claremont M, Major JE, *et al*. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res* 2007 ; 35 : D721–D726.
10. Pearl J. *Causality: models, reasoning and inference*. New York : Cambridge University Press, 2009.

NOUVELLE

Aldéhydes et anémie de Fanconi

L'ennemi de l'intérieur

Frédéric P.M. Langevin, Juan I. Garaycoechea, Gerry P. Crossan, Ketan J. Patel

Medical Research Council, Laboratory of Molecular Biology, Francis Crick avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, Royaume-Uni.
lf95@mrc-lmb.cam.ac.uk
kjp@mrc-lmb.cam.ac.uk

> La survie d'un organisme multicellulaire est étroitement liée au maintien de l'homéostasie dans l'ensemble des tissus constituant cet organisme. Cette capacité à maintenir cet équilibre vital est assurée par des populations de cellules souches, *via* trois caractéristiques principales : (1) la possibilité de se diviser et de renouveler ce compartiment en générant de nouvelles cellules souches ; (2) l'absence de fonctions tissulaires spécifiques ; et (3) la capacité de se différencier en cellules spécialisées. La moelle osseuse, où résident les cellules souches hématopoïétiques (CSH) responsables de la production des cellules du sang, constitue un tissu modèle pour l'étude des cellules souches somatiques. En effet, l'identification de multiples marqueurs de surface et le développement d'anticorps dirigés contre ces marqueurs et utilisables en cytométrie de flux (FACS, *fluorescent activated cell sorter*), permettent de discriminer et d'isoler les divers précurseurs des cellules sanguines.

Les cellules souches sont sensibles, comme toute cellule, aux conditions de stress, en particulier aux dommages de l'ADN. Ainsi, au cours du temps, l'accumulation de mutations contribue au déclin des capacités des cellules souches. L'existence de mécanismes de réparation de l'ADN est donc essentielle au maintien des populations de cellules souches [1]. Une étude récente de notre groupe décrit le rôle d'un système de réparation de l'ADN et la nature génotoxique de métabolites endogènes dans les CSH de souris [2].

L'anémie de Fanconi, une pathologie s'accompagnant d'un défaut de réparation de l'ADN

La thématique de recherche principale de notre équipe concerne l'étude des mécanismes de réparation de l'ADN et, en particulier, celle des gènes impliqués dans l'anémie de Fanconi. Notre approche est essentiellement génétique et se base sur des modèles cellulaires et murins (souris génétiquement modifiées ou *knockout* [KO]). Du nom du pédiatre suisse Guido Fanconi qui décrivit cette maladie génétique rare au début du XX^{e} siècle [3], l'anémie de Fanconi se manifeste dès l'enfance par des malformations squelettiques, une forte prédisposition à certains types de cancers (en particulier des leucémies myéloïdes) et une aplasie médullaire progressive. Les cellules de ces patients sont très sensibles à des drogues anticancéreuses, telles que la cisplatine ou la mitomycine C. Cette sensibilité constitue la base du test de diagnostic dit de « cassures chromosomiques » [4]. En effet, en liant de façon covalente deux bases opposées (interbrins d'ADN) ou adjacentes (intrabrin d'ADN), ces drogues induisent des lésions (ponts inter- ou intrabrins, *DNA crosslinks*) de l'ADN qui bloquent la réplication et la transcription. Ces ponts interbrins ont pour conséquence une augmentation des cassures chromosomiques. Dans les cellules de patients atteints d'anémie de Fanconi, ceci est dû à un défaut de réparation de l'ADN. À ce jour, une quinzaine de gènes ont été identifiés, codant pour des protéines impliquées dans cette voie de réparation des lésions pontantes de l'ADN.

Les aldéhydes, source de dommages à l'ADN

Les cassures chromosomiques se manifestent spontanément dans les cellules de patients, en dehors de la présence de drogues comme la cisplatine, ce qui suggère l'existence d'agents génotoxiques présents de façon naturelle dans les cellules de ces patients. Notre groupe a confirmé cette hypothèse en montrant en 2011 que les aldéhydes produits par le métabolisme cellulaire (en particulier l'acétaldéhyde) étaient à l'origine de dommages à l'ADN dans les cellules de patients atteints d'anémie de Fanconi [5]. Des souris double KO pour les

PLOS Computational Biology

Formal Comment Presubmission Inquiry: Human Dominant Disease Genes are Enriched in Paralogs Originating from Whole Genome Duplication

--Manuscript Draft--

Manuscript Number:	
Full Title:	Formal Comment Presubmission Inquiry: Human Dominant Disease Genes are Enriched in Paralogs Originating from Whole Genome Duplication
Article Type:	Presubmission Inquiry
Corresponding Author:	Param Priya Singh FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Param Priya Singh
First Author Secondary Information:	
Order of Authors:	Param Priya Singh
Order of Authors Secondary Information:	
Abstract:	[Info from PLOS staff: This is a Presubmission Inquiry for a Formal Comment in response to a previously published PLOS Computational Biology article (http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003073) . Please click "View Submission" to access this Formal Comment]

Human Dominant Disease Genes are Enriched in Paralogs Originating from Whole Genome Duplication

Param Priya Singh, Séverine Affeldt and Hervé Isambert
Email: param-priya.singh@curie.fr and herve.isambert@curie.fr
CNRS-UMR168, UPMC, Institut Curie, Research Center,
26, rue d'Ulm, 75248 Paris, France

PLoS Computational Biology recently published an article by Chen, Zhao, van Noort and Bork [1] reporting that human monogenic disease (MD) genes are 1) enriched in duplicates and 2) more functionally similar to their closest paralogs, in contrast to duplicated non-disease genes, based on sequence conservation and expression profile similarity. Chen *et al.* then proposed that human MD genes have frequently functionally redundant paralogs that can mask the phenotypic effects of deleterious mutations.

We would like to point out here two lines of evidence which appear more relevant to explain this surprising enrichment in duplicates of human disease genes. The first evidence is that human gene duplicates should be distinguished depending whether they originate from small scale duplication (SSD) or from the two rounds of whole genome duplication (WGD) that occurred in early vertebrates some 500 million years ago. In fact, as shown quantitatively below using Chen *et al.*'s dataset, human MD genes are actually depleted, not enriched, in SSD duplicates, whereas they are clearly enriched in WGD duplicates, when compared to non-disease genes. This opposite retention pattern cannot be explained by a selection mechanism independent of the SSD or WGD origin of MD gene duplicates. The second line of evidence concerns the mode of inheritance of human MDs, which provides a more stringent criterion than sequence conservation or co-expression profile to assess the likelihood of functional compensation by paralogs of MD genes. In particular, the recessiveness of a human disease is expected to be a prerequisite for functional compensation by a paralog gene. Indeed, autosomal dominant MDs are unlikely to experience significant functional compensation from a different locus, since even a perfectly functional allele is unable to mask the deleterious phenotypic effects of a dominant allelic mutant on the same heterozygote locus.

We first address the difference between SSD duplicates and WGD duplicates, also called "ohnologs" after Susumu Ohno's early "2R hypothesis" [2] which has now been firmly established [3]. The importance of distinguishing between SSD and WGD duplicates in the human genome has already been reported in a number of papers [4, 5, 6] including our own [7]. As shown in Figure 1A, human genes tend to partition into three main gene categories with respect to duplicates: those with WGD but no SSD duplicates (about 28%), those with SSD but no WGD duplicates (about 41%) and singletons without WGD nor SSD duplicates (about 24%), while human genes with both WGD and SSD duplicates are relatively rare (about 7%). Gene families enriched either in WGD or SSD duplicates also correspond to distinct functional classes [4, 5], with WGD genes frequently involved in signaling, regulation and development, whereas SSD genes are typically implicated in different functions such as antigen processing, immune response and metabolism.

In addition, human diseases have been shown to be enriched in WGD duplicates, while being significantly depleted in SSD duplicates [4, 7]. This could not be seen with Chen *et al.*'s dataset which

lumps together all gene duplicates irrespective of their WGD or SSD origin. In fact, using the same monogenic disease (MD) dataset readily confirms and extends these earlier results, as depicted in Figure 1B. MD genes are significantly enriched in ohnologs, 38.7% *vs* 28% ($p = 1.78 \times 10^{-64}$; χ^2 test), while showing at the same time a significant depletion in both SSD, 36.1% *vs* 41.2% ($p = 2.24 \times 10^{-13}$; χ^2 test) and singletons, 16.8% *vs* 24.1% ($p = 5.69 \times 10^{-34}$; χ^2 test). These results demonstrate that, although MD genes retain significantly more duplicates than singletons (Figure 1B), these duplicates are primarily enriched in ohnologs and not SSD copies, as compared to the relative WGD and SSD content of the entire human genome, Figure 1A.

To explain the global enrichment in MD gene duplicates, Chen *et al.* proposed that "functional compensation by duplication of genes masks the phenotypic effects of deleterious mutations and reduces the probability of purging the defective genes from the human population". However, we note that, while recent gene duplicates might be able to mask the phenotypic effect of recessive (*e.g.* loss-of-function) mutations, dominant (*e.g.* gain-of-function) mutations should typically lead to deleterious phenotypic effects regardless of the presence of any functionally redundant paralog at a different locus on the human genome.

In order to assess the extent of possible functional compensation on the retention of MD gene duplicates, we have used the mode of inheritance of human MDs rather than the less specific criteria of sequence conservation and expression profile similarity of MD genes. To this end, we retrieved the available information on the dominance and recessiveness of MDs from OMIM [8] and Blekhman *et al.* [9]. Manual curation yielded 621 autosomal dominant and 839 autosomal recessive MD genes after excluding sex linked genes and MD genes documented as both dominant and recessive. We then analyzed the duplication biases of recessive MDs (with possible functional compensation) and dominant MDs (with unlikely functional compensation). Figure 1C shows that recessive MDs do not present any biased retention of ohnologs, 28.7% *vs* 28% ($p = 0.49$; χ^2 test), SSD duplicates, 38.9% *vs* 41.2% ($p = 0.16$; χ^2 test) nor singletons, 24.1% *vs* 25.4% ($p = 0.38$; χ^2 test), as compared to their respective prevalence in the entire human genome, Figure 1A. These observations clearly show that the maintenance of recessive MD genes is in fact independent of their WGD, SSD or singleton status, suggesting limited effects of functional compensation by paralogs on the retention of gene duplicates associated to recessive MDs in human. By contrast, we observed, Figure 1D, that dominant MDs exhibit a strong enrichment in ohnologs, 45.6% *vs* 28% ($p = 1.87 \times 10^{-22}$; χ^2 test) with concomitant depletions in both SSD, 29.6% *vs* 41.2% ($p = 4.51 \times 10^{-9}$; χ^2 test), and singletons, 13.8% *vs* 24.1% ($p = 2.22 \times 10^{-9}$; χ^2 test). This is unlikely to result from a functional compensation by paralogs due to the molecular genetics of dominant MDs, as discussed above.

So, what could be the evolutionary mechanism behind the enhanced retention of WGD duplicates and depletion of SSD duplicates and singletons associated to MDs in human? In [7], we proposed a population genetics model based on the observation that a major difference between SSD and WGD scenarios concerns the timing of fixation of gene duplicates. It is well known that the SSD scenario starts with a gene duplication in the genome of a single individual, which subsequently needs to spread through the entire population to reach fixation. By contrast, the WGD scenario entails an initial fixation of duplicated gene pairs in the genome of all individuals in the small population arising through WGD. This is because WGD typically induces a speciation event due to the ploidy incompatibility of the post-WGD individuals with the rest of the pre-WGD population. This population genetics model for the fixation of SSD *versus* WGD duplicates then predicts that the

enhanced retention of “dangerous” ohnologs prone to dominant deleterious mutations is a direct consequence of purifying selection in post-WGD population, as most surviving individuals retain (non-deleterious) functional copies of their ohnologs prone to dominant deleterious mutations, while ohnologs prone to recessive deleterious mutations are more readily eliminated through loss-of-function mutations (see [7] for further details).

All in all, it appears that MD genes have preferentially retained WGD rather than SSD duplicates, as compared to non-disease genes. Yet, only dominant MD genes exhibit a clear enrichment in WGD duplicates, while the retention of recessive MD genes, which might in principle experience functional compensation from paralogs, is in fact independent of their WGD, SSD or singleton status. These results cannot be explained by the functional compensation hypothesis proposed in [1]. They are, however, consistent with a population genetics model taking into account the initial fixation of ohnologs through WGD-induced speciation and the ensuing purifying selection in post-WGD populations [7].

References

1. Chen WH, Zhao XM, van Noort V, Bork P (2013) Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. PLoS Comput Biol 9(5): e1003073.
2. Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. Hereditas 59(1): 169–187
3. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 19;453(7198): 1064-71
4. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA 18;107(20): 9270-4.
5. Huminiecki L, Heldin CH (2010) 2R and remodeling of vertebrate signal transduction engine. BMC Biol 13: 8-146.
6. Dickerson JE, Robertson DL (2012) On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol Biol Evol 29(1): 61-9.
7. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep 2(5): 1387-98.
8. McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80: 588–604.
9. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, et al. (2008) Natural selection on genes that underlie human disease susceptibility. Curr Biol 18: 883–889.

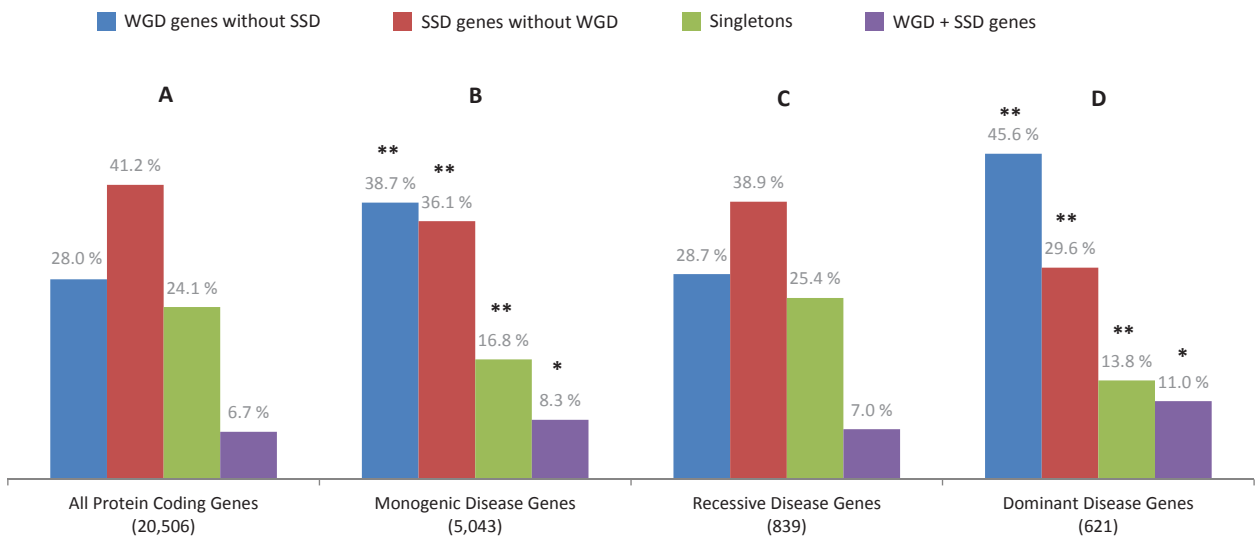

Figure 1. Distributions of WGD, SSD and singletons in (A) the whole human genome, (B) monogenic disease (MD) genes [1], (C) recessive MD genes and (D) dominant MD genes. (**) corresponds to highly significant deviations ($P < 10^{-8}$, χ^2 test) and (*) to significant deviations ($P < 10^{-4}$, χ^2 test) from the references in (A). Note that recessive MD genes (C) do not show any significant deviations in WGD, SSD nor singleton contents ($P > 10^{-1}$, χ^2 test).

List of Figures

2.1 Gene duplication by unequal crossing over . . . 8
2.2 Gene duplication by retroposition . . . 9
2.3 Whole genome duplications on the tree of life . . . 11
2.4 Whole genome duplications in the plant kingdoms . . . 12

3.1 Dosage imbalance in macromolecular complexes . . . 16

4.1 Evolution after WGD and identification of ohnologs . . . 19

6.1 Flowchart of the algorithm to identify ohnologs . . . 26
6.2 Schematic tree for the paleopolyploid and outgroup organisms with duplication nodes from Ensembl . . . 27
6.3 Identification of Macro-synteny . . . 29
6.4 The calculation of P_i . . . 31
6.5 Schematic species tree of the teleost fishes with sequenced genomes . . . 35

7.1 Mechanism of Autoinhibition . . . 41

8.1 The Mediation Diagram . . . 50

9.1 Venn diagram of distribution of human ohnologs *wrt* outgroups . . . 57
9.2 Comparisons of *P-value* distribution from original and randomized genomes . . . 58
9.3 *P-value* distribution of Makino & McLysaght ohno pairs . . . 59
9.4 *P-value* distribution of Huminiecki & Heldin ohno pairs . . . 60
9.5 Human ohnologs identified by all the three criteria . . . 61
9.6 Human ohnologs identified by intermediate/relaxed and only relaxed criteria . . . 62
9.7 Search Page on the *OHNOLOGS* server . . . 64
9.8 Ohnolog Family Page on the *OHNOLOGS* server . . . 65

10.1 Prevalence of Human Ohnologs in Gene Classes Prone to Deleterious Mutations 70
10.2 Prevalence of high confidence ohnologs in disease genes . . . 72
10.3 Prevalence of Mouse and Rat disease genes in ohnologs . . . 73
10.4 Retention of SSD/CNV duplicates in the gene classes prone to deleterious mutations . . . 74
10.5 Retention of SSD duplicates from sequence comparison in the gene classes prone to deleterious mutations . . . 75
10.6 Duplication of "dangerous" genes on evolutionary nodes leading to humans . . . 78
10.7 Ka/Ks distribution between human-amphioxus ortholog pairs . . . 80

11.1 Enrichment of dosage balanced genes in human ohnologs . . . 84
11.2 Enrichment of genes associated to complexes in human disease genes . . . 87

11.3 % Ohnologs in the highly expressed genes in the human genome. 88
11.4 Expression level distribution for ohnologs and non-ohnologs 88

12.1 Indirect effect of deleterious mutations on ohnolog retention 93

13.1 Evolutionary trends of duplicated genes following SSD or WGD 106

List of Tables

6.1 Number of protein coding genes, orthologs and paralogs for analyzed vertebrate (A) and invertebrate (B) genomes . . . 28

7.1 Cancer Databases & the Details of Obtaining Cancer Genes . . . 38
7.2 Details of 32 permanent & Transient Complexes from [Zanivan et al., 2007] . . . 44
7.3 Summary of gene counts in different categories . . . 46

8.1 Association table of binary variables to calculate direct and indirect effects . . . 51
8.2 Interpretation of Direct & Indirect Effects . . . 52

9.1 Number of **Human Ohnologs** Identified by Outgroup and Self Comparison without any filter on *P-value* and duplication time . . . 56
9.2 Individual ohnologs, pairs and families for different criteria in the human genome . . . 62
9.3 Individual ohnologs, pairs and families for all the three criteria in five vertebrate genomes 63
9.4 Individual ohnologs, pairs and families **from the 2R-WGD** for all the three criteria in Medaka, Stickleback, Tetraodon and Zebrafish genomes . . . 66
9.5 Individual ohnologs, pairs and families **from the 3R-WGD** for all the three criteria in Medaka, Stickleback, Tetraodon and Zebrafish genomes . . . 67

10.1 Dangerous gene retention in ohnologs and SSD for different human primary tumors from COSMIC . . . 77
10.2 Analysis of sequence conservation with respect to all invertebrate outgroups . . . 79

11.1 **Low retention of ohnologs from the strict criteria in permanent complexes** . . . 85

12.1 Summary of mediation analysis for dosage balance & deleterious mutation susceptibility 94
12.2 Summary of mediation analysis for essentiality & deleterious mutation susceptibility . . 99
12.3 Summary of mediation analysis for expression level & deleterious mutation susceptibility 100
12.4 Summary of mediation analysis for low Ka/Ks ratios from *Amphioxus* & deleterious mutation susceptibility . . . 102
12.5 Summary of mediation analysis for high Ka/Ks ratios from *Amphioxus* & deleterious mutation susceptibility . . . 103

Bibliography

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. (2002). Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.*, 31(1):100–105. (Cited on Page 20)

Amemiya, C. T., Alföldi, J., Lee, A. P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., et al. (2013). The african coelacanth genome provides insights into tetrapod evolution. *Nature*, 496(7445):311–316. (Cited on Page 36)

Amemiya, C. T., Powers, T. P., Prohaska, S. J., Grimwood, J., Schmutz, J., Dickson, M., Miyake, T., Schoenborn, M. A., Myers, R. M., Ruddle, F. H., and Stadler, P. F. (2010). Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc. Natl. Acad. Sci. U.S.A.*, 107(8):3622–3627. (Cited on Page 36)

Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y., et al. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282(5394):1711. (Cited on Page 11, 35, 66)

Annilo, T., Chen, Z.-Q., Shulenin, S., Costantino, J., Thomas, L., Lou, H., Stefanov, S., and Dean, M. (2006). Evolution of the vertebrate abc gene family: analysis of gene birth and death. *Genomics*, 88(1):1–11. (Cited on Page 18)

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29. (Cited on Page 42)

Assis, R. and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in drosophila. *Proc. Natl. Acad. Sci. USA*, pages 17409–14. (Cited on Page 14)

Aury, J., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116):171–178. (Cited on Page 16, 83, 84, 111)

Bailey, J. A., Liu, G., and Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.*, 73(4):823–834. (Cited on Page 8)

Baker, N. E. (2011). Cell competition. *Curr. Biol.*, 21(1):R11–R15. (Cited on Page 82)

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, 51(6):1173–1182. (Cited on Page 91)

Baxevanis, A. D. (2001). The Molecular Biology Database Collection: an updated compilation of biological database resources. *Nucleic Acids Res.*, 29(1):1–10. (Cited on Page 38)

Berry, L. W., Westlund, B., and Schedl, T. (1997). Germ-line tumor formation caused by activation of glp-1, a caenorhabditis elegans member of the notch family of receptors. *Development*, 124(4):925–936. (Cited on Page 21)

Birchler, J. and Veitia, R. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist*, 186(1):54–62. (Cited on Page 16, 42)

Birchler, J. A., Bhadra, U., Bhadra, M. P., and Auger, D. L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev. Biol.*, 234(2):275–288. (Cited on Page 15, 83, 85)

Blekhman, R., Man, O., Herrmann, L., Boyko, A., Indap, A., Kosiol, C., Bustamante, C., Teshima, K., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.*, 18(12):883–889. (Cited on Page 40, 69)

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology*, 7(5):R43. (Cited on Page 15)

Bodemann, B. and White, M. (2008). Ral gtpases and cancer: linchpin support of the tumorigenic platform. *Nature Reviews Cancer*, 8(2):133–140. (Cited on Page 41)

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K., Vogelstein, B., and Nowak, M. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA*, 107(43):18545. (Cited on Page 39)

Braasch, I., Volff, J.-N., and Schartl, M. (2009). The endothelin system: evolution of vertebrate-specific ligand–receptor interactions by three rounds of genome duplication. *Mol. Biol. Evol.*, 26(4):783–799. (Cited on Page 18)

Bridges, C. B. (1936). THE BAR 'gene' a duplication. *Science*, 83(2148):210–211. (Cited on Page 7)

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.*, 23(9):1808–1816. (Cited on Page 18, 78, 88, 101)

Byrne, K. and Wolfe, K. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3):1341. (Cited on Page 14)

Cai, J., Borenstein, E., Chen, R., and Petrov, D. (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol. Evol.*, 1:131. (Cited on Page 21, 69)

Capasso, L. (2005). Antiquity of cancer. *International journal of cancer*, 113(1):2–13. (Cited on Page 114)

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Ireland, A., Lomax, J., Carbon, S., Mungall, C., et al. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289. (Cited on Page 43)

Casewell, N. R., Wagstaff, S. C., Harrison, R. A., Renjifo, C., and Wüster, W. (2011). Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Mol. Biol. Evol.*, 28(9):2637–2649. (Cited on Page 14)

Chen, J. S., Hung, W. S., Chan, H. H., Tsai, S. J., and Sun, H. S. (2013a). In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics*, 29(4):420–427. (Cited on Page 38)

Chen, W.-H., Minguez, P., Lercher, M. J., and Bork, P. (2012). OGEE: an online gene essentiality database. *Nucleic acids research*, 40(D1):D901–D906. (Cited on Page 44, 81)

Chen, W.-H., Zhao, X.-M., van Noort, V., and Bork, P. (2013b). Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS computational biology*, 9(5):e1003073. (Cited on Page 15, 40, 85, 111)

Ciocan, C. M., Moore, J. D., and Rotchell, J. M. (2006). The role of *RAS* gene in the development of haemic neoplasia in *Mytilus trossulus*. *Marine environmental research*, 62:S147–S150. (Cited on Page 21)

Conant, G. C. and Wagner, A. (2004). Duplicate genes and robustness to transient gene knockdowns in caenorhabditis elegans. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1534):89–96. (Cited on Page 14)

David, A. and Zimmerman, M. (2010a). Cancer: an old disease, a new disease or something in between? *Nature Reviews Cancer*, 10(10):728–733. (Cited on Page 114)

David, A. and Zimmerman, M. (2010b). Cancer is an ancient disease? *Nature Reviews Cancer*, 11(1):76–76. (Cited on Page 114)

David, L., Blum, S., Feldman, M., Lavi, U., and Hillel, J. (2003). Recent duplication of the common carp (cyprinus carpio l.) genome as revealed by analyses of microsatellite loci. *Mol. Biol. Evol.*, 20(9):1425–1434. (Cited on Page 11)

Davis, J. and Petrov, D. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS biology*, 2(3):e55. (Cited on Page 88, 101)

Davis, J. and Petrov, D. (2005). Do disparate mechanisms of duplication add similar genes to the genome? *Trends in Genetics*, 21(10):548–551. (Cited on Page 15, 71, 83)

de Souza, F. S., Bumaschny, V. F., Low, M. J., and Rubinstein, M. (2005). Subfunctionalization of expression and peptide domains following the ancient duplication of the proopiomelanocortin gene in teleost fishes. *Mol. Biol. Evol.*, 22(12):2417–2427. (Cited on Page 14)

Dehal, P. and Boore, J. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10):e314. (Cited on Page 17, 20, 28, 32, 55, 114)

Denu, J. and Dixon, J. (1998). Protein tyrosine phosphatases: mechanisms of catalysis and regulation. *Current opinion in chemical biology*, 2(5):633–641. (Cited on Page 41)

Des Marais, D. and Rausher, M. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–765. (Cited on Page 14, 113)

Dickerson, J. and Robertson, D. (2011). On the origins of mendelian disease genes in man: The impact of gene duplication. *Mol. Biol. Evol.* (Cited on Page 15, 21, 113)

Dittmar, K. and Liberles, D. (2011). *Evolution after gene duplication*. John Wiley & Sons. (Cited on Page 13)

Domazet-Lošo, T. and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.*, 25(12):2699–2707. (Cited on Page 21)

Drea, S. C., Lao, N. T., Wolfe, K. H., and Kavanagh, T. A. (2006). Gene duplication, exon gain and neofunctionalization of oep16-related genes in land plants. *The Plant Journal*, 46(5):723–735. (Cited on Page 14)

Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N., et al. (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of arabidopsis. *Mol. Biol. Evol.*, 23(2):469–478. (Cited on Page 14)

Dufresne, F. and Hebert, P. D. (1995). Polyploidy and clonal diversity in an arctic cladoceran. *Heredity*, 75(1):45–53. (Cited on Page 11)

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Airey, M., Anagnostopoulos, A., Babiuk, R., Baldarelli, R., et al. (2012). The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40(Database issue):D881–886. (Cited on Page 43, 73, 81)

Eudeline, B., Allen Jr, S. K., and Guo, X. (2000). Optimization of tetraploid induction in pacific oysters, *Crassostrea gigas*, using first polar body as a natural indicator. *Aquaculture*, 187(1):73–84. (Cited on Page 11)

Evangelisti, A. M. and Conant, G. C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biol Evol*, 2:826–834. (Cited on Page 85)

Evlampiev, K. and Isambert, H. (2007). Modeling protein network evolution under genome duplication and domain shuffling. *BMC systems biology*, 1(1):49. (Cited on Page 83, 111)

Evlampiev, K. and Isambert, H. (2008). Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc. Natl. Acad. Sci. USA*, 105(29):9863. (Cited on Page 111)

Faltas, B. (2010). Cancer is an ancient disease: the case for better palaeoepidemiological and molecular studies. *Nature Reviews Cancer*, 11(1):76–76. (Cited on Page 114)

Fares, M. A., Keane, O. M., Toft, C., Carretero-Paulet, L., and Jones, G. W. (2013). The roles of whole-genome and small-scale duplications in the functional specialization of saccharomyces cerevisiae genes. *PLoS genetics*, 9(1):e1003176. (Cited on Page 15, 71)

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. *Nucleic acids research*, 41(D1):D48–D55. (Cited on Page 25, 38)

Forbes, S., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J., Futreal, P., and Stratton, M. (2008). The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics*. (Cited on Page 38, 76)

Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531. (Cited on Page 14)

Francino, M. P. (2005). An adaptive radiation model for the origin of new gene functions. *Nat. Genet.*, 37(6):573–578. (Cited on Page 14)

Freeling, M. (2008). The evolutionary position of subfunctionalization, downgraded. *Genome Dyn*, 4:25–40. (Cited on Page 113)

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual review of plant biology*, 60:433–453. (Cited on Page 111, 113)

Freeling, M. and Thomas, B. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome research*, 16(7):805. (Cited on Page 15)

Furney, S., Albà, M., and López-Bigas, N. (2006). Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC genomics*, 7(1):165. (Cited on Page 69)

Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. (2004). A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183. (Cited on Page 38)

Gallardo, M., Gonzalez, C., and Cebrian, I. (2006). Molecular cytogenetics and allotetraploidy in the red vizcacha rat, *Tympanoctomys barrerae* (rodentia, octodontidae). *Genomics*, 88(2):214–221. (Cited on Page 12)

Gambi, M., Ramella, L., Sella, G., Protto, P., and Aldieri, E. (1997). Variation in genome size in benthic polychaetes: systematic and ecological relationships. *Journal of the Marine Biological Association of the United Kingdom*, 77(04):1045–1057. (Cited on Page 11)

Gibson, T. and Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in Genetics*, 14(2):46–49. (Cited on Page 22)

Gibson, T. A. and Goldberg, D. S. (2009). Questioning the ubiquity of neofunctionalization. *PLoS computational biology*, 5(1):e1000252. (Cited on Page 13)

Gillis, W. Q., St John, J., Bowerman, B., and Schneider, S. Q. (2009). Whole genome duplications and expansion of the vertebrate gata transcription factor gene family. *BMC evolutionary biology*, 9(1):207. (Cited on Page 18)

Gong, X., Wu, R., Zhang, Y., Zhao, W., Cheng, L., Gu, Y., Zhang, L., Wang, J., Zhu, J., and Guo, Z. (2010). Extracting consistent knowledge from highly inconsistent cancer gene data sources. *BMC bioinformatics*, 11(1):76. (Cited on Page 37)

Gout, J., Duret, L., and Kahn, D. (2009). Differential retention of metabolic genes following whole-genome duplication. *Mol. Biol. Evol.*, 26(5):1067–1072. (Cited on Page 45, 86, 100)

Gout, J., Kahn, D., and Duret, L. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics*, 6(5):e1000944. (Cited on Page 45, 83, 86, 100)

Gregory, T. R., Hebert, P. D., and Kolasa, J. (2000). Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. *Heredity*, 84(2):201–208. (Cited on Page 11)

Grozeva, S., Kuznetsova, V., and Nokkala, S. (2004). Patterns of chromosome banding in four nabid species (heteroptera, cimicomorpha, nabidae) with high chromosome number karyotypes. *Hereditas*, 140(2):99–104. (Cited on Page 11)

Gu, X. (2003). Evolution of duplicate genes versus genetic robustness against null mutations. *Trends in Genetics*, 19(7):354–356. (Cited on Page 14)

Gu, Z., Steinmetz, L., Gu, X., Scharfe, C., Davis, R., and Li, W. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66. (Cited on Page 14, 15, 99)

Guan, Y., Dunham, M. J., and Troyanskaya, O. G. (2007). Functional analysis of gene duplications in saccharomyces cerevisiae. *Genetics*, 175(2):933–943. (Cited on Page 15, 99)

Hakes, L., Pinney, J., Lovell, S., Oliver, S., and Robertson, D. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome biology*, 8(10):R209. (Cited on Page 15, 71, 83, 99)

Haldane, J. B. S. (1932). *The causes of evolution*. Princeton University Press. (Cited on Page 7)

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 33(Database issue):D514–517. (Cited on Page 38)

Hastings, P., Lupski, J., Rosenberg, S., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564. (Cited on Page 8, 10)

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., et al. (2012). A census of human soluble protein complexes. *Cell*, 150(5):1068–1081. (Cited on Page 42, 43)

He, X. and Zhang, J. (2006). Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.*, 23(1):144. (Cited on Page 99)

Higgins, M., Claremont, M., Major, J., Sander, C., and Lash, A. (2006). CancerGenes: a gene selection resource for cancer genome projects. *Nucleic acids research*, 35(suppl 1):D721. (Cited on Page 39)

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A., Delany, M. E., et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716. (Cited on Page 63)

Holland, L., Albalat, R., Azumi, K., Benito-Gutiérrez, È., Blow, M., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L., et al. (2008). The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research*, 18(7):1100–1111. (Cited on Page 17)

Holland, L. Z. and Short, S. (2008). Gene duplication, co-option and recruitment during the origin of the vertebrate brain from the invertebrate chordate brain. *Brain, behavior and evolution*, 72(2):91–105. (Cited on Page 17)

Holland, P. W., Garcia-Fernàndez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development*, 1994(Supplement):125–133. (Cited on Page 17)

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. (Cited on Page 67)

Huang, S., Tian, H., Chen, Z., Yu, T., and Xu, A. (2010). The evolution of vertebrate tetraspanins: gene loss, retention, and massive positive selection after whole genome duplications. *BMC evolutionary biology*, 10(1):306. (Cited on Page 18)

Hubbard, S. (2002). Autoinhibitory mechanisms in receptor tyrosine kinases. *Front Biosci*, 7:d330–d340. (Cited on Page 41)

Hubbard, S. (2004). Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat. Rev. Mol. Cell Biol.*, 5:465. (Cited on Page 41)

Hubbard, S., Mohammadi, M., and Schlessinger, J. (1998). Autoregulatory mechanisms in protein-tyrosine kinases. *Journal of Biological Chemistry*, 273(20):11987. (Cited on Page 41)

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.*, 256(1346):119–124. (Cited on Page 14, 106)

Hughes, M. K. and Hughes, A. L. (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.*, 10(6):1360–1369. (Cited on Page 11)

Huminiecki, L. and Heldin, C. (2010). 2R and remodeling of vertebrate signal transduction engine. *BMC biology*, 8(1):146. (Cited on Page 15, 55, 58, 60, 61, 62, 111)

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. *PLoS biology*, 2(7):e206. (Cited on Page 8)

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957. (Cited on Page 11, 18, 35, 66, 67, 114)

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467. (Cited on Page 18)

Jiang, W.-K., Liu, Y.-L., Xia, E.-H., and Gao, L.-Z. (2013). Prevalent role of gene features in determining evolutionary fates of wgd duplicated genes in flowering plants. *Plant physiology*. (Cited on Page 88, 101)

Jiao, Y., Wickett, N., Ayyampalayam, S., Chanderbali, A., Landherr, L., Ralph, P., Tomsho, L., Hu, Y., Liang, H., Soltis, P., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*. (Cited on Page 11)

Jin, J., Xie, X., Chen, C., Park, J. G., Stark, C., James, D. A., Olhovsky, M., Linding, R., Mao, Y., and Pawson, T. (2009). Eukaryotic protein domains as functional units of cellular evolution. *Science signaling*, 2(98):ra76. (Cited on Page 14)

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, 10(1):19–31. (Cited on Page 9, 10)

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the caenorhabditis elegans genome using rnai. *Nature*, 421(6920):231–237. (Cited on Page 15)

Kellis, M., Birren, B., and Lander, E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624. (Cited on Page 14, 18, 85, 114)

Kersey, P. J., Staines, D. M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J. C., Hughes, D. S., Keenan, S., Kerhornou, A., Koscielny, G., et al. (2012). Ensembl genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic acids research*, 40(D1):D91–D97. (Cited on Page 25)

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., et al. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37(Database issue):D767–772. (Cited on Page 42)

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. (2003). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *The EMBO journal*, 23(1):234–243. (Cited on Page 10)

Kuraku, S. and Meyer, A. (2009). The evolution and maintenance of hox gene clusters in vertebrates and the teleost-specific genome duplication. *Int J Dev Biol*, 53(5-6):765–773. (Cited on Page 17)

LaFave, M. C. and Sekelsky, J. (2009). Mitotic recombination: why? when? how? where? *PLoS genetics*, 5(3):e1000411. (Cited on Page 8)

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948. (Cited on Page 47)

Laulederkind, S. J., Hayman, G. T., Wang, S. J., Smith, J. R., Lowry, T. F., Nigam, R., Petri, V., de Pons, J., Dwinell, M. R., Shimoyama, M., Munzenmaier, D. H., Worthey, E. A., and Jacob, H. J. (2013). The Rat Genome Database 2013–data, tools and users. *Brief. Bioinformatics*, 14(4):520–526. (Cited on Page 46, 73)

Levinson, G. and Gutman, G. (1987). Slipped-strand mispairing: a major mechanism for dna sequence evolution. *Mol. Biol. Evol.*, 4(3):203–221. (Cited on Page 10)

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J. K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, 34(Database issue):D572–580. (Cited on Page 60)

Liang, H. and Fernandez, A. (2008). Evolutionary constraints imposed by gene dosage balance. *Front Biosci*, 13:4373–4378. (Cited on Page 42)

Liang, H. and Li, W. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends in Genetics*, 23(8):375–378. (Cited on Page 15, 99)

Liao, B. and Zhang, J. (2007). Mouse duplicate genes are as essential as singletons. *Trends in Genetics*, 23(8):378–381. (Cited on Page 15, 99)

Liao, B. and Zhang, J. (2008). Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl. Acad. Sci. USA*, 105(19):6987. (Cited on Page 44)

Lin, Y., Hwang, J., and Li, W. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome. *Gene*, 387(1-2):109–117. (Cited on Page 85, 99)

Lynch, M. and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151. (Cited on Page 13, 74)

Lynch, M. and Conery, J. (2003). The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, 3(1):35–44. (Cited on Page 74)

Lynch, M. and Force, A. (2000a). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459. (Cited on Page 14, 106)

Lynch, M. and Force, A. G. (2000b). The origin of interspecific genomic incompatibility via gene duplication. *The American Naturalist*, 156(6):590–605. (Cited on Page 105, 113)

Mable, B. (2004). 'why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society*, 82(4):453–466. (Cited on Page 10)

MacCarthy, T. and Bergman, A. (2007). The limits of subfunctionalization. *BMC Evol. Biol.*, 7:213. (Cited on Page 113)

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA*, 102(15):5454. (Cited on Page 15, 16, 83, 111)

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 39(Database issue):D52–57. (Cited on Page 38)

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 2011:bar009. (Cited on Page 38)

Makino, T., Hokamp, K., and McLysaght, A. (2009). The complex relationship of gene duplication and essentiality. *Trends in Genetics*, 25(4):152–155. (Cited on Page 15, 43, 99)

Makino, T. and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. USA*, 107(20):9270. (Cited on Page 15, 16, 20, 28, 32, 45, 55, 58, 59, 60, 61, 62, 71, 83, 84, 91, 111, 113)

Malaguti, G., Singh, P. P., and Isambert, H. (2013). On the retention of gene duplicates prone to dominant deleterious mutations. *Submitted: Theoretical Population Biology*. (Cited on Page 112, 113)

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., and Otto, S. P. (2011). Recently formed polyploid plants diversify at lower rates. *Science*, 333(6047):1257. (Cited on Page 10)

McLysaght, A., Hokamp, K., and Wolfe, K. (2002). Extensive genomic duplication during early chordate evolution. *Nat. Genet.*, 31(2):200–204. (Cited on Page 17)

Moran, J. V., DeBerardinis, R. J., and Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science*, 283(5407):1530–1534. (Cited on Page 10)

Muller, H. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist*, 59(663):346–353. (Cited on Page 10)

Nadeau, J. and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147(3):1259–1266. (Cited on Page 112)

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *science*, 338(6105):384–387. (Cited on Page 14, 113)

Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. *Nature*, 221(5175):40–42. (Cited on Page 7)

Noonan, J. P., Grimwood, J., Danke, J., Schmutz, J., Dickson, M., Amemiya, C. T., and Myers, R. M. (2004). Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res.*, 14(12):2397–2405. (Cited on Page 36)

Ohno, S. (1970). *Evolution by gene duplication*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. (Cited on Page 7, 10, 11, 13)

Ohno, S., Wolf, U., and Atkin, N. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187. (Cited on Page 7, 10, 11, 17)

Okada, K. and Asai, K. (2008). Expansion of signaling genes for adaptive immune system evolution in early vertebrates. *BMC genomics*, 9(1):218. (Cited on Page 17)

Opazo, J. C., Butts, G. T., Nery, M. F., Storz, J. F., and Hoffmann, F. G. (2012). Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Mol. Biol. Evol.* (Cited on Page 18)

Papp, B., Pál, C., and Hurst, L. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197. (Cited on Page 15, 16, 83, 84, 91)

Pearl, J. (2001). Direct and Indirect Effects. *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA: Morgan Kaufmann*, pages 411–420. (Cited on Page 49, 50, 51, 52, 91, 95, 96, 97, 98)

Pearl, J. (2009). *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition. (Cited on Page 49, 50, 52)

Pearl, J. (2012a). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *In Causality: Statistical Perspectives and Applications (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). John Wiley and Sons, Ltd, Chichester, UK*, pages 151–179. (Cited on Page 49, 50, 51, 52, 91, 95, 96, 97, 98)

Pearl, J. (2012b). The causal mediation formula–a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436. (Cited on Page 49, 50, 51, 52, 95, 96, 97, 98)

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., et al. (2012). The gencode pseudogene resource. *Genome Biol*, 13:R51. (Cited on Page 9, 10)

Podder, S. and Ghosh, T. C. (2011). Insights into the molecular correlates modulating functional compensation between monogenic and polygenic disease gene duplicates in human. *Genomics*, 97(4):200–204. (Cited on Page 40)

Pufall, M. A. and Graves, B. J. (2002). Autoinhibitory domains: modular effectors of cellular regulation. *Annu. Rev. Cell Dev. Biol.*, 18:421–462. (Cited on Page 41)

Putnam, N. H., Butts, T., Ferrier, D. E., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Yu, A. T. J.-K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071. (Cited on Page 17, 18, 20, 21, 25, 28, 55, 56, 114)

Qian, W. and Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics*, 179(4):2319. (Cited on Page 84)

Robert, J. (2010). Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates. *Developmental & Comparative Immunology*, 34(9):915–925. (Cited on Page 21)

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155. (Cited on Page 91)

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H. W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, 38(Database issue):497–501. (Cited on Page 42)

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010:baq020. (Cited on Page 38)

Santarius, T., Shipley, J., Brewer, D., Stratton, M., and Cooper, C. (2010). A census of amplified and overexpressed human cancer genes. *Nature Reviews Cancer*, 10(1):59–64. (Cited on Page 38)

Santini, F., Harmon, L., Carnevale, G., and Alfaro, M. (2009). Did genome duplication drive the origin of teleosts? a comparative study of diversification in ray-finned fishes. *BMC evolutionary biology*, 9(1):194. (Cited on Page 66)

Sémon, M. and Wolfe, K. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in xenopus laevis. *Proc. Natl. Acad. Sci. USA*, 105(24):8333. (Cited on Page 78, 88, 101, 113)

Sémon, M. and Wolfe, K. H. (2007). Consequences of genome duplication. *Current opinion in genetics & development*, 17(6):505–512. (Cited on Page 15, 111)

Semyonov, J., Park, J.-I., Chang, C. L., and Hsu, S. Y. T. (2008). Gpcr genes are preferentially retained after whole genome duplication. *PLoS One*, 3(4):e1903. (Cited on Page 18)

Seoighe, C. and Wolfe, K. (1999). Yeast genome evolution in the post-genome era. *Current opinion in microbiology*, 2(5):548–554. (Cited on Page 45, 86, 100)

Siegel, R., Naishadham, D., and Jemal, A. (2013). Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, 63(1):11–30. (Cited on Page 114)

Silva, J. M., Marran, K., Parker, J. S., Silva, J., Golding, M., Schlabach, M. R., Elledge, S. J., Hannon, G. J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, 319(5863):617–620. (Cited on Page 44, 81)

Simillion, C., Janssens, K., Sterck, L., and Van de Peer, Y. (2008). i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, 24(1):127–128. (Cited on Page 58)

Singh, P. P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., and Isambert, H. (2012). On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell reports*. (Cited on Page 47, 58, 59, 62, 89, 101, 111, 112, 113)

Soshnikova, N., Dewaele, R., Janvier, P., Krumlauf, R., and Duboule, D. (2013). Duplications of hox gene clusters and the emergence of vertebrates. *Dev. Biol.* (Cited on Page 17)

Steinke, D., Hoegg, S., Brinkmann, H., and Meyer, A. (2006). Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC biology*, 4(1):16. (Cited on Page 18)

Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2012). Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol. Phylogenet. Evol.* (Cited on Page 18)

Sturtevant, A. H. (1925). The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics*, 10(2):117–147. (Cited on Page 7)

Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101(16):6062. (Cited on Page 45)

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160. (Cited on Page 34)

Taylor, J., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research*, 13(3):382. (Cited on Page 66)

Taylor, J. H., Woods, P. S., and Hughes, W. L. (1957). The organization and duplication of chromosomes as revealed by autoradiographic studies using tritium-labeled thymidinee. *Proc. Natl. Acad. Sci. USA*, 43(1):122–128. (Cited on Page 7)

Tinti, M., Johnson, C., Toth, R., Ferrier, D., and MacKintosh, C. (2012). Evolution of signal multiplexing by 14-3-3-binding 2R-ohnologue protein families in the vertebrates. *Open Biology*, 2(7). (Cited on Page 18)

Tirosh, I. and Barkai, N. (2007). Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome biology*, 8(4):R50. (Cited on Page 14)

Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome research*, 13(12):2559. (Cited on Page 9)

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725–732. (Cited on Page 17)

Veitia, R. (2002). Exploring the etiology of haploinsufficiency. *Bioessays*, 24(2):175–184. (Cited on Page 15, 83)

Veitia, R. (2003). Nonlinear effects in macromolecular assembly and dosage sensitivity. *Journal of Theoretical Biology*, 220(1):19–25. (Cited on Page 15)

Veitia, R. (2009). On gene dosage balance in protein complexes: A comment on Semple JI, Vavouri T, Lehner B. A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC systems biology*, 3(1):16. (Cited on Page 42)

Vilella, A., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327. (Cited on Page 27, 76)

Volinia, S., Mascellani, N., Marchesini, J., Veronese, A., Ormondroyd, E., Alder, H., Palatini, J., Negrini, M., and Croce, C. (2008). Genome wide identification of recessive cancer genes by combinatorial mutation analysis. *PLoS One*, 3(10):e3380. (Cited on Page 38)

Wada, H. (2010). Origin and genetic evolution of the vertebrate skeleton. *Zoological science*, 27(2):119–123. (Cited on Page 17)

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010a). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, 8(1):77–80. (Cited on Page 47)

Wang, Y., Zhang, T., and Wang, W. (2010b). An old disease, a new disease or something in between: evidence from china. *Nature Reviews Cancer*, 11(1):76–76. (Cited on Page 114)

Wolfe, K. (2000). Robustness—it's not where you think it is. *Nat. Genet.*, 25(1):3–4. (Cited on Page 17)

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C., Haase, J., Janes, J., Huss 3rd, J., et al. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10(11):R130. (Cited on Page 45, 87)

Yang, Y. and Fu, L. M. (2003). TSGDB: a database system for tumor suppressor genes. *Bioinformatics*, 19(17):2311–2312. (Cited on Page 38)

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17(1):32–43. (Cited on Page 46, 47)

Zanivan, S., Cascone, I., Peyron, C., Molineris, I., Marchio, S., Caselle, M., and Bussolino, F. (2007). A new computational approach to analyze human protein complexes and predict novel protein interactions. *Genome Biology*, 8(12):R256. (Cited on Page 43, 44, 141)

Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011). Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res.*, 21(12):2114–2128. (Cited on Page 85)

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., and Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.*, 41(7):849–853. (Cited on Page 10)

Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., and Scherer, S. W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.*, 115(3-4):205–214. (Cited on Page 45)

Zhao, M., Sun, J., and Zhao, Z. (2013). TSgene: a web resource for tumor suppressor genes. *Nucleic acids research*, 41(D1):D970–D976. (Cited on Page 38)

Expansion des familles de gènes impliquées dans des maladies par duplication du génome chez les premiers vertébrés

L'expansion au cours de l'évolution de familles de gènes impliquées dans les cancers et d'autres maladies génétiques graves est surprenante du point de vue de la sélection naturelle. Dans cette thèse, nous avons montré que des familles de gènes sujettes à des mutations délétères dans le génome humain se sont principalement agrandies par rétention de gènes "ohnologues" issus de deux duplications globales du génome (GGD) datant de l'origine des vertébrés à mâchoires. En utilisant une méthode d'inférence avancée, nous avons aussi démontré que la rétention de nombreux ohnologues soupçonnés d'être susceptibles aux équilibres de dosage d'expression était en fait plus directement liée à leur sensibilité aux mutations délétères. Cette rétention priviligiée d'ohnologues "dangereux", définis comme sujets à des mutations délétères dominantes, semble être une conséquence des évênements de spéciation provoqués par ces GGD et la sélection de purification qui a suivi dans les espèces post-GGD. Nous avons également développé une approche quantitative pour identifier les ohnologues dans le génome des vertébrés. Ces ohnologues sont facilement accessibles à partir d'un serveur Web. Nos résultats soulignent l' importance de la sélection non adaptative induite par GGD dans l'émergence de la complexité des vertébrés, tout en rationalisant, d'un point de vue évolutif, l'extension des familles de gènes fréquemment impliquées dans les maladies génétiques et les cancers. Les ohnologues identifiés par notre approche ouvrent également la voie à de nouvelles analyses de génomique fonctionnelle distinguant l'origine des gènes dupliqués.

Expansion of disease gene families by whole genome duplication in early vertebrates

The emergence and evolutionary expansion of gene families implicated in cancers and other severe genetic diseases is an evolutionary oddity from a natural selection perspective. In this thesis, we have shown that gene families prone to deleterious mutations in the human genome have been preferentially expanded by the retention of "ohnolog" genes from two rounds of whole-genome duplication (WGD) dating back from the onset of jawed vertebrates. Using advanced inference analysis, we have further demonstrated that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations. This enhanced retention of "dangerous" ohnologs, defined as prone to autosomal-dominant deleterious mutations, is shown to be a consequence of WGD-induced speciation and the ensuing purifying selection in post-WGD species. We have also developed a statistical approach to identify ohnologs in vertebrate genomes with high confidence. These ohnologs can be easily accessed from a web server. Our findings highlight the importance of WGD-induced non-adaptive selection for the emergence of vertebrate complexity, while rationalizing, from an evolutionary perspective, the expansion of gene families frequently implicated in genetic disorders and cancers. The high confidence ohnologs identified by our approach will also pave the way for novel functional genomic analyses distinguishing gene duplicates according to their origin.